Lung and Colon Cancer Detection Using Custom CNN on Histopathological Images

Varun Magotra
Department of Computer Engineering
Sardar Patel Institute of Technology
Mumbai, India
varun.magotra@spit.ac.in

Ebrahim Gabajiwala
Department of Computer Science
Sardar Patel Institute of Technology
Mumbai, India
ebrahim.gabajiwala@spit.ac.in

Ebrahim Hirani
Department of Computer Science
Sardar Patel Institute of Technology
Mumbai, India
ebrahim.hirani@spit.ac.in

Jainam Jain
Department of Computer Science
Sardar Patel Institute of Technology
Mumbai, India
jainam.jain@spit.ac.in

Sudhir N Dhage
Department of Computer Science
Sardar Patel Institute of Technology
Mumbai, India
sudhir_dhage@spit.ac.in

***Abstract*—Lungs and Colon cancer have been two of the most significant causes of cancer related deaths, worldwide. So, detecting and accurately diagnosing cancer is of paramount importance. CNNs have been giving promising results in various classification, identification and image processing tasks since their inception. Due to this, CNNs have been prominently used in the healthcare sector for medical image analysis, image segmentation, etc. Using these automated systems has helped medical personnel not only to correctly detect a disease but has also made the diagnosis process swift and economical. However, bettering the accuracy of the diagnosis is challenging. In this study, we proposed a CNN architecture to classify the Lung and Colon Cancer Histopathological Images. The model is trained and tested for 3 different modes of configurations(Lung tissue Classification, Colon Tissue Classification and Lung+Colon Tissue Classification). To compare the performance of our model, we used the pre-trained EfficientNet B0 on the same dataset with the same 3 configurations. Our model,with 595K parameters, is able to get accuracy exceeding 99% on all the three configurations and its performance is at par with the EfficientNet B0, which has more than 5 million parameters. This light model may be easily inferred on low performance devices, eliminating the long time required to process these samples in laboratories.**

***Index Terms*—CNN, EfficientNet B0, Image Processing**

I. Introduction

Cancer is a plague that has been haunting our world since the very beginning.The first documented case of Cancer hails from ancient Egypt in 1500 BC. The details were recorded on papyrus, documenting eight cases of tumors occurring on the breast. In 2020, there were 19.3 million new cancer cases and 10 million deaths reported due to cancer. Out of the 19.3 million new cancer cases, lung and colon cancer were responsible for 11.4% and 5.8% of them respectively. Lung and colon cancer accounted for 18% and 5.8%, respectively, of the 10 million cancer deaths in 2020 [1] . In India, cumulative risk of developing Lung and bronchus cancer is 1 in 101 and for colon cancer is 1 in 298 which is very alarming [2].

Medical science has made a lot of progress in the last 2 decades but still we are far from having a perfect cure for cancer and this exacerbates the situation because of the smaller recovery rate from it. Various methods and processes are used to diagnose cancers. Some patients are not wealthy enough to afford laboratory techniques and expert supervision. The resources used for the establishment of these labs can be used to solve other problems like medication and treatment of the ailment.

In order to overcome the cost of lab and expert personnel we must adopt an automated way to diagnose cancer. The advent of application of deep learning and machine learning in the area of medical sciences has been proven to be very economical. In the last 3-4 years, new deep learning architectures and state of the art computer vision models have proved to be quite effective in the analysis of medical images, bolstering medical science. These models and techniques are showing promising results in disease classification, segmentation [3 ,4].

Here, we propose a deep learning model to supplant expensive equipment and expert personnel. Rest of the paper is structured as follows: In Section II, we briefly discuss related works. Section III Dataset and PreProcessing. Section IV Methodology, Section V Result and Evaluation, Section VI Conclusion.

II. Related Studies

The first application of artificial intelligence in the medical field through the invention of MYCIN in 1977[5]. Since then, machine learning and deep learning approaches have evolved rapidly and their implication in the medical sector has become wide spread. Convolutional neural networks(CNN'S) are widely used for the classification, segmentation and detection of anomalies from medical images [6]. CNN has the ability to capture contextual details in independent parts of images, without being affected by minor changes in the image as a whole.They also require less number of parameters compared to ordinary ANNs for the same input size to generate a specific output size. Because of this convoluted neural networks have

played a vital role in image processing and computer vision tasks[3,7].

Various machine learning algorithms have been developed and used for identifying the cancerous cells. After the introduction of Artificial intelligence in medical science, the first task was to tackle cancer disease as it was one of the leading cause of death worldwide. Teramoto et. al., in their research they developed a Deep Convolution Neural Network (DCNN), which consists of three convolution layers,three pooling layers and two fully connected layers.While evaluation they trained the model using microscopic images which were first cropped and resampled to obtain images images with the resolution of 256 x 256 pixels and to gain the perfect fit the dataset was augmented via rotation, flipping and filtering.Their proposed algorithm was able to classify 71% images correctly[8].

In 2020, Mangal et. al. were able to achieve an accuracy of 97% on lung cancer and 96% on colon cancer using a shallow neural network for classification. A traditional convolutional neural network architecture was used with convolutional layers, pooling layers and dropout layers followed by a flatten and fully connected layer for classification. Their experiments demonstrated an improvement over traditional machine learning and deep learning architectures.[9]

In 2021, Masud et.al., in their study used digital image processing techniques along with deep learning models to classify 5 different types of lungs and colon tissues. Masud et.al., used two 2DWavelet, 2DFourier algorithms for feature extraction from the images and then after concatenating the features from the two algorithms, they classified them using the CNN's. The authors were able to identify the cancerous cells for both the lungs and colon with the accuracy of 96.33

In 2014, the idea of transfer learning for pre-trained models was introduced. The idea behind the transfer learning is to use the pre-trained weights from huge models trained on large datasets, and then use these weights in models built for smaller datasets. Transfer learning techniques were proposed which freeze the initial parameters that capture the features of the input and allow the last classification layers to adapt and change according to the dataset. Yosinski et. al. explored the use of transfer learning in their study and demonstrated its effectiveness in improving performance. M. Hussain et al. performed a comparative study in which the re-trained Inception V3 model with transfer learning gave better results than the model trained on the CIFAR-10 images from scratch. [11, 12]

Abbas et al. use ResNet-18, ResNet-34, ResNet-50, ResNet 101, VGG 19 and Alex-Net, pre-trained on ImageNet, to classify the 15000 lung images divided in 3 classes of the Lung and Colon Cancer Histopathological Image Dataset dataset. They achieve accuracies of 98.8%, 99.18%, 99.6%, 99.8%, 98.93% and 97.26%, respectively. [13]

These methods have given effectual results for image classification. In our paper, we are proposing a custom CNN architecture for identifying the lungs and colon tissues and comparing it with the state of the art model-EfficientNet[14].

III. Dataset and PreProcessing

For our study, we used the LC25000 dataset [15]. This dataset contains Histopathological images classified into 5 types of Lung and Colon tissue as shown in figure 1 and described in table I:

TABLE I: Class Description

Serial No.	*Types of Tissue*	*Class Name*	*Count*
1	Lung benign tissue	lung_n	5000
2	Lung adenocarcinoma	$\text{lung}_a ca$	5000
3	Lung squamous cell carcinoma	$\text{lung}_s cc$	5000
4	Colon adenocarcinoma	$\text{colon}_a ca$	5000
5	Colon benign	colon_n	5000

Fig. 1: Example of each class from the Lung Colon Dataset

This dataset contains around 25000 square images of size 768 pixels. This Dataset is derived from a dataset containing 750 images for lung tissues and 500 images of colon tissues through image augmentation. We normalize the pixel values of all the images in our dataset to values between 0 to 1. This reduces the computational requirements demanded by higher values, and also requires lower parameter values, allowing for higher learning rates. After normalizing the images, we rescaled the images to the size of 256x 256 pixels. We split the dataset into training, validation and testing in the ratio of 80:10:10 respectively.

The lung tissue images and colon tissue images are different folders, allowing us to study both of these separately. We also combine them to create a model that is able to classify all the 5 classes at a time.

IV. Methodology

In our study, we proposed a custom CNN architecture for Lung Colon Tissue classification to identify which tissues can cause cancer and which tissues are benign. This model architecture is shown in figure 2.

Our model consists of 6 components:

- **Convolution**: We use 7 convolutional layers in our model. Convolutional layers are the most widely used components in computer vision systems. An odd sized square filter (3x3 in our model) is applied to the image with a specific stride (1 in our case). We progressively double the number of filters applied in every layer. The advantages these layers provide are due to their filters, that are independently applied to each part of the image, helping it determine features irrespective of their positions in the image. Convolutional layers are also resource efficient compared to fully connected neural networks as they require less number of parameters for the same sizes of

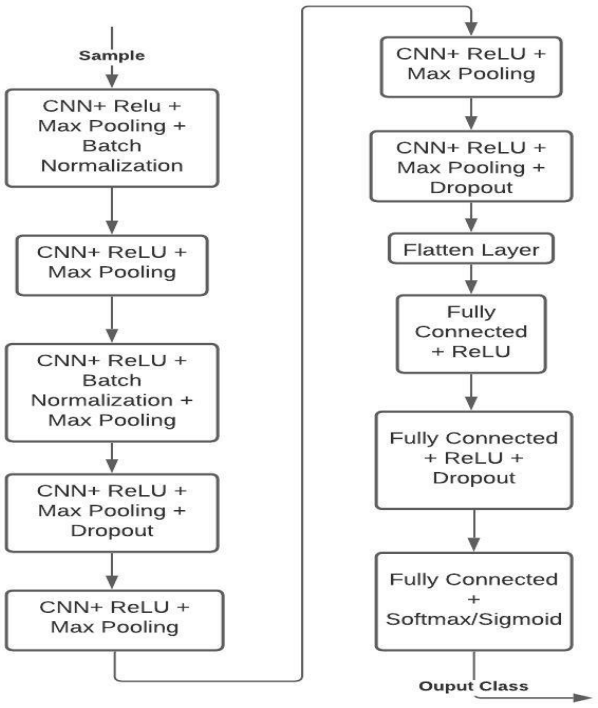


Fig. 2: Custom CNN Architecture

inputs and outputs required. ReLU activation function is applied after each layer to introduce non linearity. ReLU stands for Rectified Linear Activation Function which outputs the input for positive values but returns 0 for other values as shown in figure 3

- **Pooling**: Pooling layer is responsible for reducing the spatial resolution of the convolved features.This layer is added to the architecture in order to reduce the computational power by reducing the dimensions.In addition it is used to exclude key rotating and dynamic features that exist, thus maintaining a successful model training process. There are mainly two types of pooling layers: Max Pooling and Average Pooling. We use one pooling layer after each convolution layer with pooling size of 2,padding set to valid.All the pooling layers in our CNN model use Max Pooling Operations.In max pooling approach , it returns the maximum value from the portion of image covered by the Kernel.
- **Batch Normalization layer**: Batch Normalization is a technique used to standardized the inputs from the layers. Batch Normalization helps to use higher learning rates and accelerates the training process by providing some regularization which in turn reduces the generalization errors[16]. In our model's architecture, we used the Batch Normalization two times (as shown in figure 2) so that the generalization errors would be low and also our model would train more quickly and efficiently.
- **DropOut layer**: We use dropout layers at three positions, one before the 5th Convolutional Layer, one before the Flatten layer and one before the final Fully Connected layer (as shown in figure 1). This layer helps prevent overfitting in Supervised Learning tasks. Units are dropped randomly along with their connections, preventing them from co-adapting. Each unit is dropped with the specified probability, independent of the other units [17].
- **Flatten**: We apply this layer after we are finished applying all the convolutional layers and have received all the feature maps from the model. This converts the multidimensional output of the convolutional layers to a single array, which can be then used to apply classification techniques.
- **Fully Connected Layers**: We apply 3 fully connected layers in our model that carry out the classification task. Softmax activation is applied to the final layer for models giving 5 and 3 classes as output, and sigmoid activation is applied for the model giving 2 classes.

So, depending upon the classification, only the last layer of our model is changed but the rest of the architecture remains the same.

To compare the accuracy and efficiency of our model, we compare its performance with the State of the art model- EfficientNet by performing transfer learning[11,14]. EfficientNets scale the model size using certain coefficients which are used to set the width, depth and resolution sizes.

After Initializing these models, we trained them in 3 different configurations:

1. Lung Tissues Classification (3 classes)
2. Colon Tissues Classification (2 classes)
3. Lung + Colon Tissues Classification (5 classes)

V. Results and Discussion

Depending on the model and the configuration for which we are doing the classification, training time per epochs, number of epochs, accuracy, loss varied differently as mentioned in table II.

Configuration	Model	Accuracy		Loss	
		Training	Validation	Training	Validation
Colon	Custom CNN Architecture	99.93	100	0.0034	0.0011
	EfficientNetB0	98.71	100	0.0362	0.0073
Lung	Custom CNN Architecture	99.85	99.33	0.004227	0.016995
	EfficientNetB0	97.725	99.2667	0.060102	0.028142
Lung+Colon	Custom CNN Architecture	99.667	99.598	0.0091	0.0187
	EfficientNetB0	98.487	99.357	0.0483	0.0272

TABLE II: Accuracy and Loss Comparison of Custom CNN and EfficientNetB0

The CNN architecture that we proposed in our study is able to get higher accuracy than the existing models and frameworks that have been proposed for the LC25000 dataset [15]. After training the CNN architecture, we trained EfficientNet using transfer learning on the same dataset. For EfficientNet, the accuracy comes out to be 99.267 on lungTissue Classification, 100 on ColonTissue and 99.357on Lung+Colon Tissues while, the accuracy of CNN comes out to be 99.333 on lungTissue Classification, 100 on ColonTissue and 99.598 on Lung+Colon Tissues on the validation dataset. The training

accuracy and loss for CNN and Efficient Net on the three configurations can be seen in the figure 3,4,5.

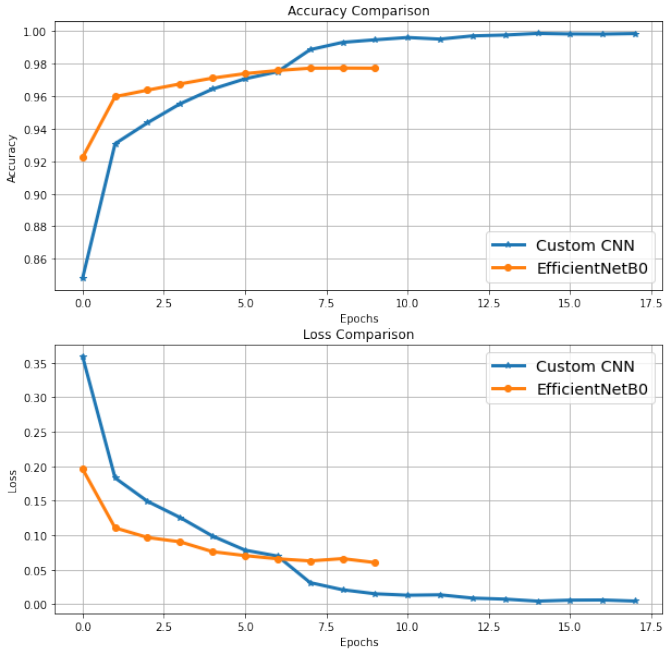


Fig. 3: Accuracy and Loss Comparison on Lung Tissues Classification

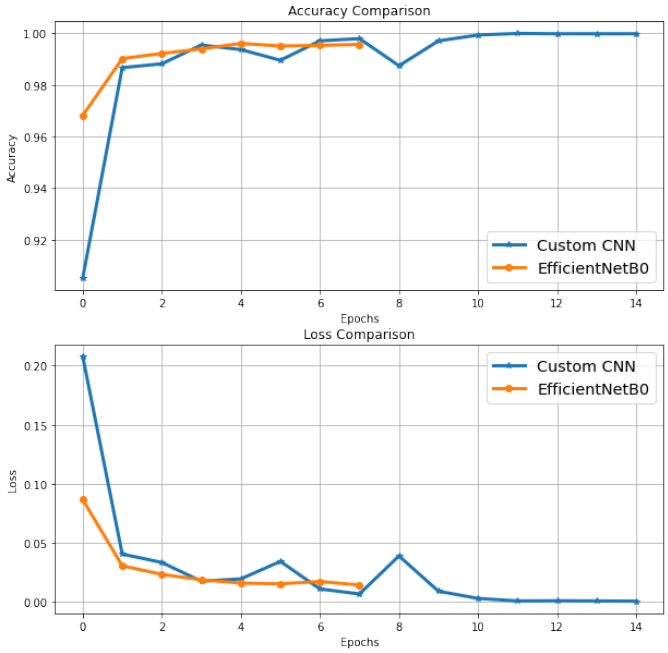


Fig. 4: Accuracy and Loss Comparison on Colon Tissues Classification

As the result indicates , our CNN model is able to correctly classify each of the tissues more accurately and precisely as compared to EfficientNet. Moreover, our custom CNN has

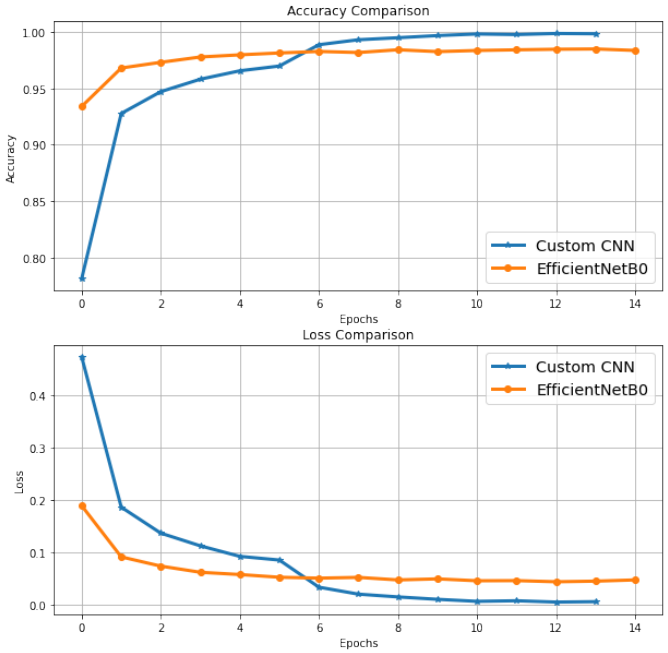


Fig. 5: Accuracy and Loss Comparison on Lung+Colon Tissues Classification

just over 595K parameters compared to more than 5 million parameters in EfficientNet-B0.

After training the models on 3 different modes of our dataset. We evaluated and tested the model on the test set using the following parameters : Accuracy, Precision, Recall and F1-score.

- **Accuracy:** Accuracy is ratio of number of correctly classified samples to the total number of samples. It helps us to know how accurately one's model is in terms of getting the right results.

$$Accuracy = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

- **Precision:** Precision is defined as the closeness of two or more measurements with each other. It records the true positive of the model and tells us what fraction of that predicted true values are actually true.

$$Precision = \frac{\text{True Positive}}{\text{True Positive +False Positive}}$$

- **Recall:** It is the rate at which the model gives a correct positive value for actual positive values.

$$Recall = \frac{\text{True Positive}}{\text{True Positive +False Negative}}$$

- **F1-Score:** The F1-score is a way of combining the precision and recall of the model, and it is defined as the harmonic mean of the model's precision and recall.

$$F_1 = 2 * \frac{\text{Precision * Recall}}{\text{Precision + Recall}}$$

where

As shown in table III,IV, the performance of our CNN model is at par or better than the Efficient Net. It can be clearly observed that Accuracy, Precision, Recall, F1- Score for all the three modes of Categorization is close to 1 or equal to 1.

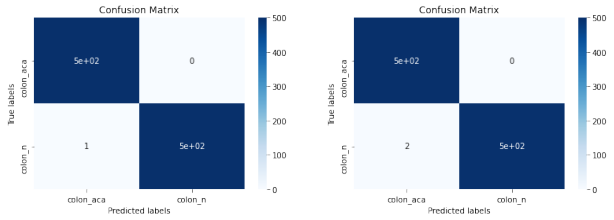
Configuration	Model	Test Accuracy
Colon	Custom CNN Architecture	0.9990
	EfficientNetB0	0.9980
Lung	Custom CNN Architecture	0.9927
	EfficientNetB0	0.9833
Lung +Colon	Custom CNN Architecture	0.9928
	EfficientNetB0	0.9928

TABLE III: Accuracy Comparison of Models on Test Data

Data	Model	Class	Precision	Recall	F1-score
Colon	Custom CNN	colon_n	1.00	1.00	1.00
		colon_aca	1.00	1.00	1.00
	EfficientNetB0	colon_n	1.00	1.00	1.00
		colon_aca	1.00	1.00	1.00
Lung	Custom CNN	lung_aca	0.99	0.99	0.99
		lung_scc	1.00	1.00	1.00
		lung_n	0.99	0.99	0.99
	EfficientNetB0	lung_aca	0.99	0.97	0.98
		lung_scc	1.00	1.00	1.00
		lung_n	0.97	0.99	0.98
L+C	Custom CNN	colon_n	1.00	1.00	1.00
		lung_aca	1.00	1.00	1.00
		lung_scc	0.99	0.97	0.98
		lung_n	1.00	1.00	1.00
		colon_aca	0.97	0.99	0.98
	EfficientNetB0	colon_n	1.00	1.00	1.00
		lung_aca	1.00	1.00	1.00
		lung_scc	0.98	0.99	0.98
		lung_n	1.00	1.00	1.00
		colon_aca	0.99	0.98	0.98

TABLE IV: Precision, Recall, F1-score Comparison

As we see, in the smaller sample size of 500 samples for each class for the test set, both the models give really high values with little to no errors in outputs. And we visualized the classification results for the test data in the form of Confusion Matrix as shown in figure 10,11,12.

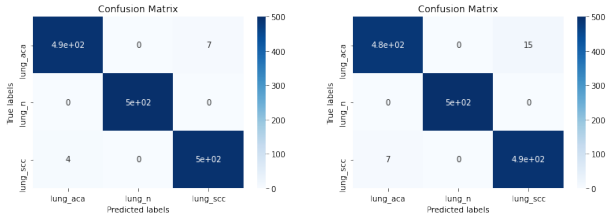


(a) Custom CNN (b) EfficientNet B0

Fig. 6: Confusion Matrix For Colon Tissue Classification on Unseen Test Data

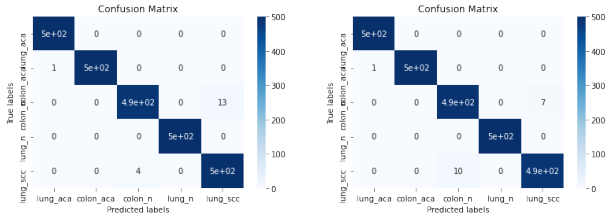
VI. Conclusion

Classification of cancer tissues is a meticulous task because of its minute details that need to be taken care of while classifying it in the laboratory. So, to reduce those struggles and reduce the human errors on the final result we propose a deep learning model which gives accuracy of about 99% in most cases and almost 100% in some cases. Our model has only a fraction of the number of parameters of that of the state of the art models and runs faster during inference as well due to the simplicity of the model. This model helps to automate cancer detection, and a pipeline can be created to get lab samples, preprocess the image, feed the image to the model and receive the test result. Our proposed CNN model has performed better than or at par with EfficientNet, which is one of the state of the art models.



(a) Custom CNN (b) EfficientNet B0

Fig. 7: Confusion Matrix For Lung Tissue Classification on Unseen Test Data



(a) Custom CNN (b) EfficientNet B0

Fig. 8: Confusion Matrix For Lung+Colon Tissue Classification on Unseen Test Data

References

[1] Hyuna Sung PhD, Jacques Ferlay MSc, ME, Rebecca L. Siegel MPH, Mathieu Laversanne MSc, Isabelle Soerjomataram MD, MSc, PhD, Ahmedin Jemal DMV, PhD, Freddie Bray BSc, MSc, PhD. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries, 2021

[2] DOI: 10.1200/GO.20.00122 JCO Global Oncology no. 6 (2020) 1063-1075. Published online July 16, 2020.

[3] Hou, L. et al. Patch-based convolutional neural network for whole slide tissue image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2424–2433 (2016).

[4] HarDNet-MSEG: A Simple Encoder-Decoder Polyp Segmentation Neural Network that Achieves over 0.9 Mean Dice and 86 FPS.

[5] Shortliffe, Edward. (1977). Mycin: A Knowledge-Based Computer Program Applied to Infectious Diseases*. Proceedings / the ... Annual Symposium on Computer Application [sic] in Medical Care. Symposium on Computer Applications in Medical Care.

[6] Yamashita R, Nishio M, Do RKG, Togashi K. Convolutional neural networks: an overview and application in radiology. Insights Imaging. 2018;9(4):611-629. doi:10.1007/s13244-018-0639-9

[7] Bengio, Y. Lecun, Yann. (1997). Convolutional Networks for Images, Speech, and Time-Series.

[8] Teramoto A, Tsukamoto T, Kiriyama Y, Fujita H. Automated Classification of Lung Cancer Types from Cytological Images Using Deep Convolutional Neural Networks. Biomed Res Int. 2017;2017:4067832. doi:10.1155/2017/4067832

[9] Sanidhya Mangal and Aanchal Chaurasia and Ayush Khajanchi. Convolution Neural Networks for diagnosing colon and lung cancer histopathological images. arXiv:2009.03878, 2020.

[10] Masud, Mehedi Sikder, Niloy Nahid, Abdullah Bairagi, Anupam AlZain, Mohammed. (2021). A Machine Learning Approach to Diagnosing Lung and Colon Cancer Using a Deep Learning-Based Classification Framework. Sensors. 21. 10.3390/s21030748.

[11] Yosinski J, Clune J, Bengio Y, and Lipson H. How transferable are features in deep neural networks? In Advances in Neural Information Processing Systems 27 (NIPS '14), NIPS Foundation, 2014.

[12] Hussain, Mahbub and Bird, Jordan J. and Faria, Diego R.A Study on CNN Transfer Learning for Image Classification,Advances in Computational Intelligence Systems. Springer International Publishing, 2019.

[13] The Histopathological Diagnosis of Adenocarcinoma Squamous Cells Carcinoma of Lungs by Artificial intelligence: A comparative study of convolutional neural networks Mohammad Ali Abbas, Syed Usama Khalid Bukhari, Asmara Syed, Syed Sajid Hussain Shah medRxiv 2020.05.02.20044602; doi: https://doi.org/10.1101/2020.05.02.20044602

[14] Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks, arXiv:1905.11946, 2020.

[15] Andrew A. Borkowski and Marilyn M. Bui and L. Brannon Thomas and Catherine P. Wilson and Lauren A. DeLand and Stephen M. Mastorides. Lung and Colon Cancer Histopathological Image Dataset (LC25000). arXiv:1912.12142, 2019.

[16] Sergey Ioffe and Christian Szegedy.Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift,arXiv:1502.03167,2015.

[17] Nitish Srivastava and Geoffrey Hinton and Alex Krizhevsky and Ilya Sutskever and Ruslan Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research, 2014.

[18] E. Hirani, V. Magotra, J. Jain and P. Bide, "Plant Disease Detection Using Deep Learning," 2021 6th International Conference for Convergence in Technology (I2CT), 2021, pp. 1-4, doi: 10.1109/I2CT51068.2021.9417910.

[19] Lakhani P, Sundaram B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology. 2017;284:574–582. doi: 10.1148/radiol.2017162326

[20] Kermany, Daniel S., et al. "Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning." Cell, vol. 172, no. 5, 2018, pp. 1122–1131.

[21] P. Dutta, P. Upadhyay, M. De and R. G. Khalkar, "Medical Image Analysis using Deep Convolutional Neural Networks: CNN Architectures and Transfer Learning," 2020 International Conference on Inventive Computation Technologies (ICICT), 2020.

[22] Alexander Selvikvåg Lundervold, Arvid Lundervold, An overview of deep learning in medical imaging focusing on MRI, Zeitschrift für Medizinische Physik, Volume 29, Issue 2, 2019, Pages 102-127, ISSN 0939-3889.

[23] Raghu, Maithra Zhang, Chiyuan Kleinberg, Jon Bengio, Samy. (2019). Transfusion: Understanding Transfer Learning with Applications to Medical Imaging.

[24] Mohammad Ali Kadampur, Sulaiman Al Riyaee, Skin cancer detection: Applying a deep learning based model driven architecture in the cloud for classifying dermal cell images, Informatics in Medicine Unlocked, Volume 18, 2020, 100282, ISSN 2352-9148

[25] Kumar, K. Venkatalakshmi, K. Krishnan, Karthikeyan. (2019). Lung Cancer Detection Using Image Segmentation by means of Various Evolutionary Algorithms. Computational and Mathematical Methods in Medicine. 2019. 1-16. 10.1155/2019/4909846.