Lung and Colon Cancer Classification and Detection Using Convolutional Neural Network Architecture MobileNet

REZUAN ISLAM (id: 20-43691-2)[1],
MD. MASHRURUL HAQUE (id: 20-43751-2)[2],
TAMANNA AKTER (id: 20-42410-1)[3],
DAVASIS SARKAR (id: 20-42429-1)[4]

[a]*Department of Computer Sciences, American International University-Bangladesh*

***Abstract*— Lung and colon are the dangerous kinds of cancer that need to be detected as soon as possible for effective treatment. Using specialists for examining histological images is both time-consuming and error-pruning. So, we have used CNN to solve this problem. CNN can identify and classify lung and colon cancer types with better accuracy in a shorter period, for determining patients' right treatment this is crucial. Normal tissue, adenocarcinoma, squamous cell carcinoma, large cell carcinoma, colon adenocarcinoma, and colon benign tissue are considered in this research work with MobileNet architecture. The Convolutional Neural networks model obtained training and validation accuracy of 95.34% and 93.67%.**

Keywords*— Deep learning, Classification, CNN, MobileNet

I. INTRODUCTION

Cancer is a primary reason for mortality on a global scale, and the most common and deadly types are colon and lung cancers. Lung cancer contributes to 18.4% of cancer-related deaths, while colon cancer accounts for 9.2% of total cancer-related deaths worldwide, as reported by the World Health Organization [2]-[3]. Approximately 17% of people encounter the concurrent manifestation of lung and colon cancer [1]. The lack of timely detection can result in a significant increase in the spread of cancer cells between the lungs and colon [4]. Effective treatment and early detection are crucial in reducing cancer mortality rates [5]. Timely diagnosis improves management, enhances the chances of recovery, and increases patient survival rates.

To detect cancerous cells and eliminate other potential conditions, various tests like imaging (x-ray, CT scan), sputum cytology, and tissue sampling (biopsy) are conducted. However, accurately diagnosing and classifying lung and colon cancers based on histopathology slides is a time-consuming process for pathologists and medical professionals. This poses a risk of misdiagnosis, leading to incorrect treatment and potential harm to patients. Our objective is to develop a solution that expedites and improves the precision of the diagnostic process.

Machine learning algorithms play a crucial role in biomedical applications, enabling the prediction and classification of various signals and images. Previous research articles have explored the use of deep learning (DL) for simultaneous classification of lung and colon cancer images. While some authors focused on lung cancer classification, others concentrated solely on colon cancer. For instance, Bukhari et al. [8] achieved the highest accuracy of 93.91% using three CNN architectures. Hatuwal et al. [9] achieved a classification result of 97.2% for histopathological images of lung cancer using a CNN. Mangal et al. [10] applied a shallow neural network. architecture and achieved 97% and 96% accuracy in classifying lung and colon cancers, respectively. Previous papers typically classified at most three different types of lung cancer or focused solely on colon cancer, and the dataset sizes were insufficient for obtaining robust results.

This article aims to utilize a Convolutional Neural Network (CNN) architecture to classify Normal, Large Cell Carcinoma, Adenocarcinoma, Squamous Cell Carcinoma, Colon Adenocarcinoma, and Colon Benign Tissue using Histopathological and CT scan images. The primary objective of this study is to propose a medical diagnostic support system for lung and colon imaging, incorporating a larger dataset and a broader range of cancer types. Additionally, we seek to enhance interpretability, and understanding of how the CNN makes decisions, and explore the use of techniques like Grad-CAM or LIME to build trust among medical professionals.

Section II provides a review of previous related works, while Section III offers a concise description of the methodology and settings employed. The research findings are presented and illustrated with plots and tables in Section IV. The paper concludes in Section V, followed by a reference section listing the cited sources.

II. RELATED WORKS

The author B. K. Hatuwal and H. C. Thapa [11] aims to use Convolutional Neural Networks for classifying lung cancer histopathological images into three categories: benign tissue, Adenocarcinoma, and squamous cell carcinoma. The authors highlight the importance of early detection for better patient recovery and the limitations of current diagnostic methods, which can be error-prone and time-consuming. The literature review section of the paper provides a brief overview of related research on lung cancer detection using various techniques, such as deep learning, image processing, and machine learning. It mentions studies that have used X-ray, CT scan images, and histopathological images for lung cancer detection. The authors highlight the low accuracy of some previous models and emphasize the use of CNNs in their work for better accuracy.

The author of this paper focuses on lung cancer segmentation, a crucial research topic in medical imaging. they discuss the significance of accurate segmentation in determining the effectiveness of anticancer drugs and texture analyses on medical images. They highlight the limitations of manual segmentation due to interobserver variability and propose the use of deep learning models for automatic segmentation to overcome this issue. The paper presents a novel approach that combines generative adversarial

networks (GANs) and transfer learning with pretrained models to address the small dataset problem in medical imaging. They generate an artificial dataset of lung nodules using GANs and use it to train pretrained models. This allows them to perform transfer learning for lung cancer segmentation without requiring manual annotation for labeling the dataset [12].

O. Ronneberger, P. Fischer, and T. Brox introduces a novel deep learning network and training strategy for image segmentation, leveraging data augmentation to optimize the use of limited annotated samples. The architecture comprises a contracting path for context capture and a symmetric expanding path for precise localization. By training on very few images, the network outperforms previous methods on the ISBI challenge for segmentation of neuronal structures and won the ISBI cell tracking challenge 2015 for microscopy images. Notably, the network demonstrates impressive speed, processing a 512x512 image in less than a second on a recent GPU. literature review contains related papers on deep learning Network and Training Strategy for Image Segmentation with Limited Annotated Samples [13].

M. Beldar, S. Rajguru, S. Suman, T. B. Patil's paper compares various algorithmic methods for early lung cancer detection from CT images. The study employs the ResNet-50 transfer learning model to potentially increase accuracy, inspired by its success in COVID-19 and breast cancer detection. The primary goal is to identify the best technique for improved lung cancer diagnosis, aiding timely treatment and better patient outcomes. A summary of relevant papers on lung cancer detection utilizing various techniques, such as image processing, deep learning, and CNN, is provided by the literature review [14].

M. Schultheiss in this paper highlights the challenges of interpreting CXR images, especially in detecting small lung nodules. Deep learning-based computer-aided diagnosis (CAD) systems have gained interest in the research community due to advancements in computing power. These systems have shown promising results in disciplines like breast cancer screening and melanoma image classification, sometimes even exceeding human reader performance. The authors mention that current state-of-the-art methods for detecting abnormalities in medical images, such as pneumothorax and mammography screening, utilize deep learning models trained with detailed annotations like box coordinates or segmentations. The Retina.Net [15].

III. PRPOSED MODEL

Our proposed system followed dataset acquisition, data formatting, CNN Model architecture, activation function, optimizer, model training, validation, and testing, described in the below sections.

A. Dataset Acquisition

We consider two datasets for our work. One is LC25000 Lung and colon histopathological image dataset[16], and another is CT-Scan images Dataset[17]. The images are divided into 6 classes which are Normal tissue, Adenocarcinoma, Squamous Cell Carcinoma, Large Cell Carcinoma, Colon adenocarcinoma, and Colon benign tissue. Each part has 5000 images except Large Cell Carcinoma. This has 187 images.

Class Distribution of the dataset

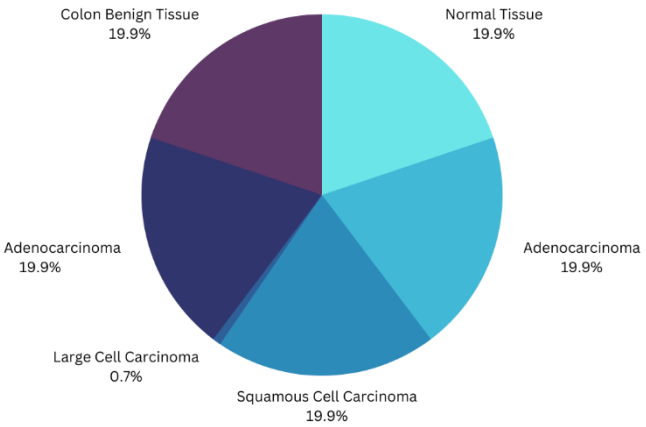

Fig-1(a): Class Distribution of the dataset

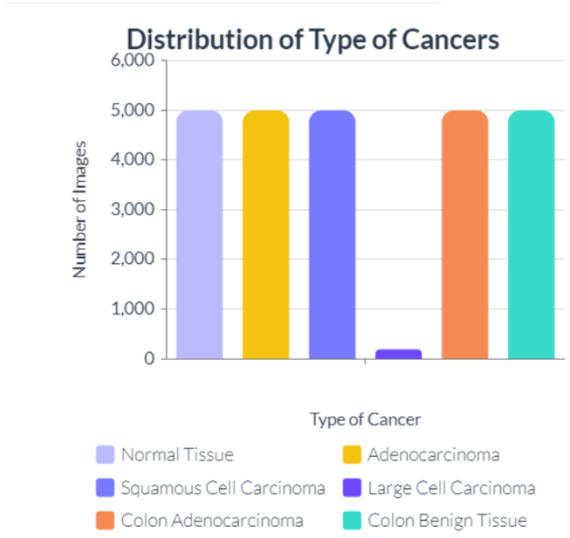

Fig-1(b): Distribution of the types of lung and colon cancer

B. Data Formatting

We have used data augmentation for our image preprocessing. In data augmentation, the images were resized in (224, 224) pixel sizes to maintain a uniform aspect ratio. All the images' shapes were converted in a range of (0,1). The batch size was set to 128 so that the neural network would not tend to over-fit. Fig. 1(a) and Fig. 1(b) are the visualizations of different types of histopathology images and their augmented images.

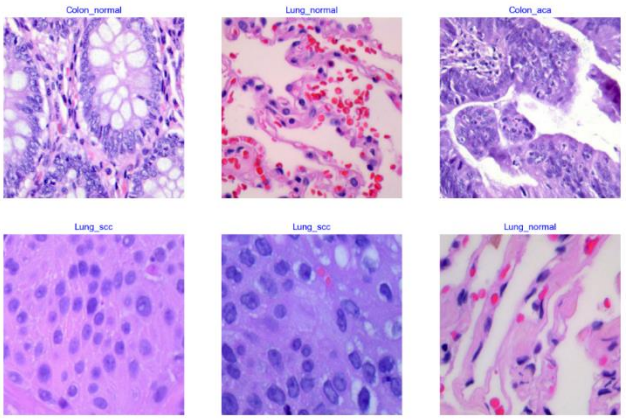

Fig-2(a): Histopathology Images of different categories

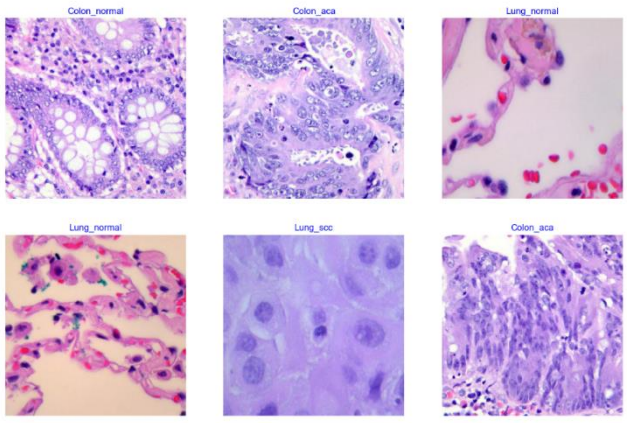

Fig-2(b): Histopathology Images of different categories

C. Model Training, Validation and Testing

Images are split in a ratio of 70:20:10 for training, validation, and testing purposes.

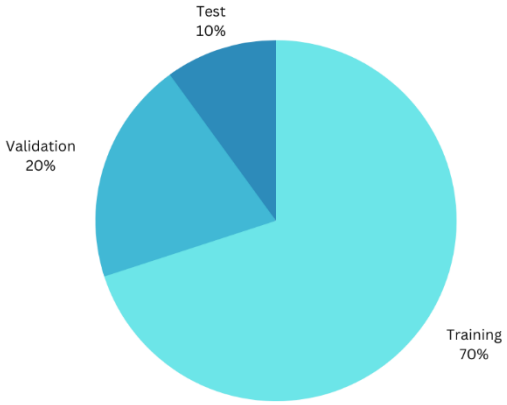

Fig-3: Images split ratio for training, validation, and testing

CNN Model Architecture: We created a model using MobileNet architecture, where we have loaded our input dataset and added a layer of 32 convolutional filters. Then added a batch normalization and passed the convolutional kernel in 7*7 average pooling layer. After that, we pooled a feature map and applied a dense layer with 256 nodes. A dropout value of 0.25 was applied to the model. A dense value of seven with the sigmoid activation function was used to obtain the class probabilities for final output classes.
Our model -

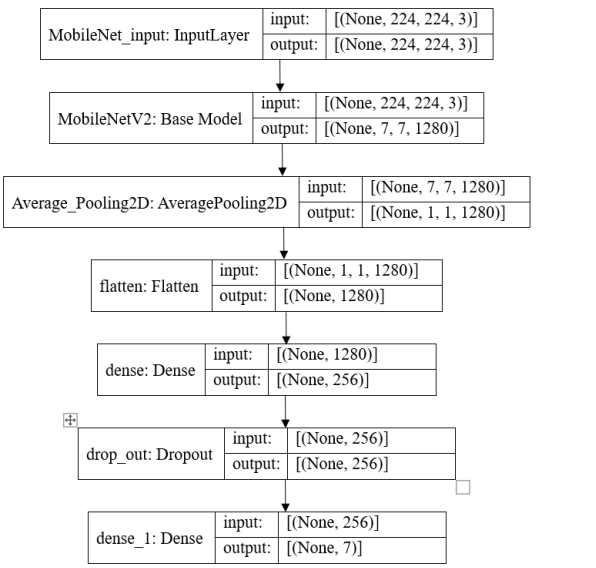

Fig-4: Proposed model

Activation functions and Optimizers: We have at first used different activation functions and optimizers. After testing those activation functions and optimizers, we decided to use the sigmoid activation function and Adam optimizer for our model.

The confusion matrix plot was used to measure the performance of the developed CNN model. Also, the metrics accuracy, precision, recall, and f1-score were calculated as below:

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN}$$

$$Precision = \frac{TP}{(TP + FP)}$$

$$Recall = \frac{TP}{(TP + FN)}$$

$$F1 - Score = \frac{2 * (Recall * Precision)}{(Recall + Precision)}$$

Where TP means true positive, FP means false positive, FN means false negative, and TN means true negative values. This trained model weights were saved into the hd5 file format and also used to test or predict the cancer type by loading the weights to the model architecture.

IV. RESULT

The images were trained for 28 epochs with batch size 138. We used different activation functions and different optimizers to evaluate the performance of each proposed model where we got the best accuracy by using the Sigmoid activation function and RMSprop optimizer. In this model we achieved a training accuracy of 95.34% and a validation accuracy of 93.61%.

Table:

Comparison of different models.

Final Layer Activation Methos	Optimizer	Training Accuracy (%)	Validation Accuracy (%)
Softmax	RMSprop	77.0	76.52
Softmax	Adam	94.11	91.92
Sigmoid	RMSprop	95.34	93.61

1 For Softmax Activation function and RMSprop Optimizer:

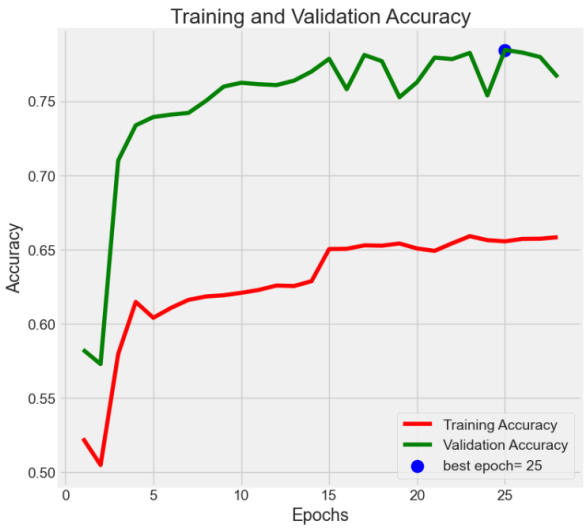

Fig-4: Plot of Model Accuracy vs. Epoch for Training and Validation Images

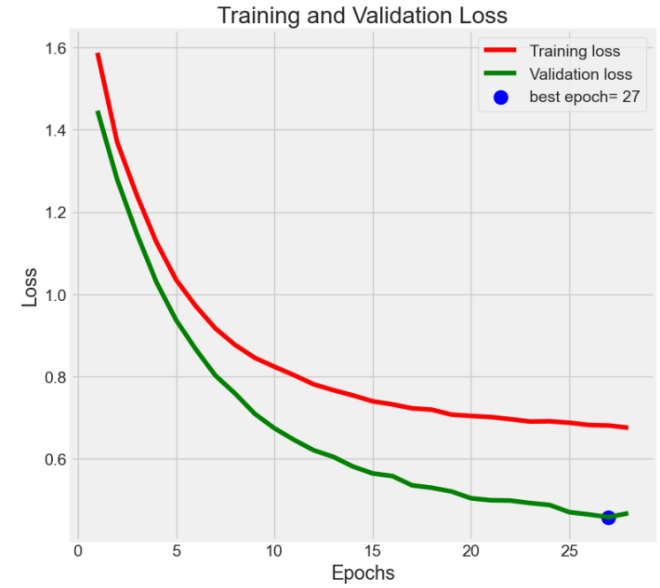

Fig-5: Plot of Model Loss vs. Epoch for Training and Validation Images

	precision	recall	f1-score	support
Colon_aca	0.50	0.04	0.07	500
Colon_normal	0.51	0.98	0.67	500
Lung_aca	0.94	0.87	0.90	500
Lung_lcc	1.00	1.00	1.00	51
Lung_normal	0.98	1.00	0.99	500
Lung_scc	0.90	0.95	0.92	500
accuracy			0.77	2551
macro avg	0.80	0.80	0.76	2551
weighted avg	0.77	0.77	0.72	2551

Table-2: Accuracy for Sigmoid activation function and RMSprop optimizer.

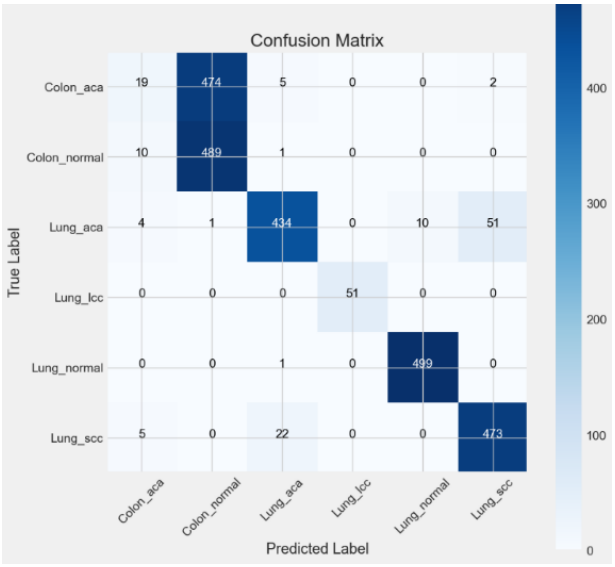

Fig-6: Confusion Matrix of Different Image Categories for Validation Images

2) Softmax Activation function and Adam Optimizer:

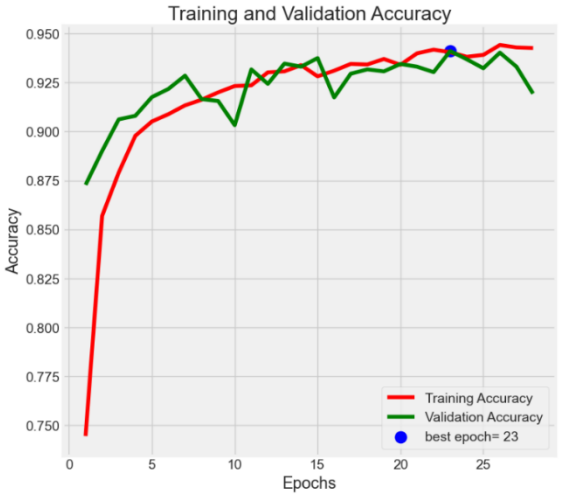

Fig-7: Plot of Model Accuracy vs. Epoch for Training and Validation Images

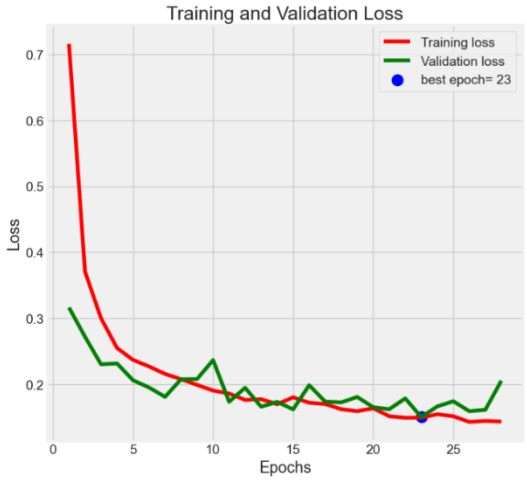

Fig-8: Plot of Model Loss vs. Epoch for Training and Validation Images

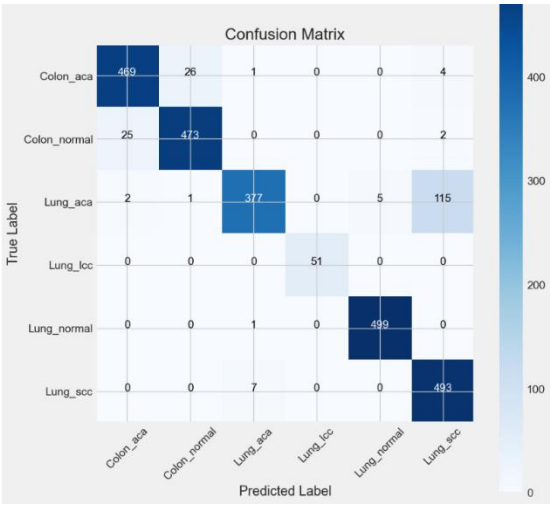

Fig-9: Confusion Matrix of Different Image Categories for Validation Images

	precision	recall	f1-score	support
Colon_aca	0.95	0.94	0.94	500
Colon_normal	0.95	0.95	0.95	500
Lung_aca	0.98	0.75	0.85	500
Lung_lcc	1.00	1.00	1.00	51
Lung_normal	0.99	1.00	0.99	500
Lung_scc	0.80	0.99	0.89	500
accuracy			0.93	2551
macro avg	0.94	0.94	0.94	2551
weighted avg	0.93	0.93	0.93	2551

Table-3: Accuracy for Softmax Activation function and Adam Optimizer.

3) For Sigmoid Activation function and RMSprop Optimizer:

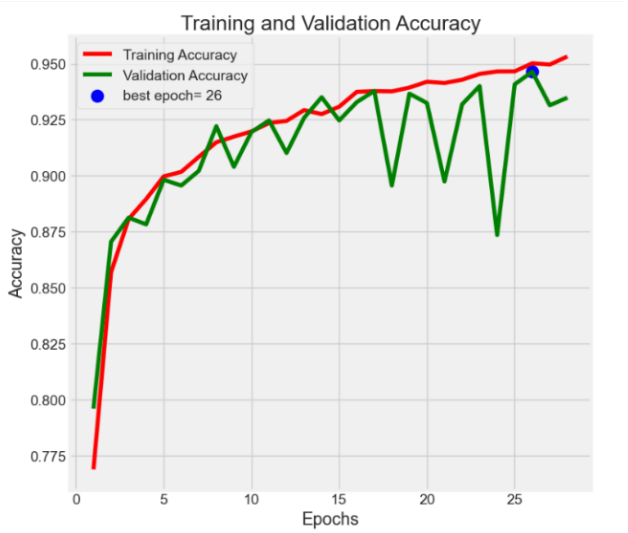

Fig-10: Plot of Model Accuracy vs. Epoch for Training and Validation Images

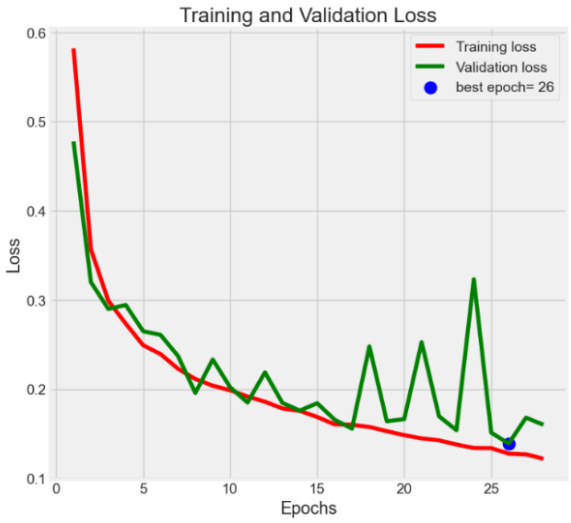

Fig-11: Plot of Model Loss vs. Epoch for Training and Validation Images

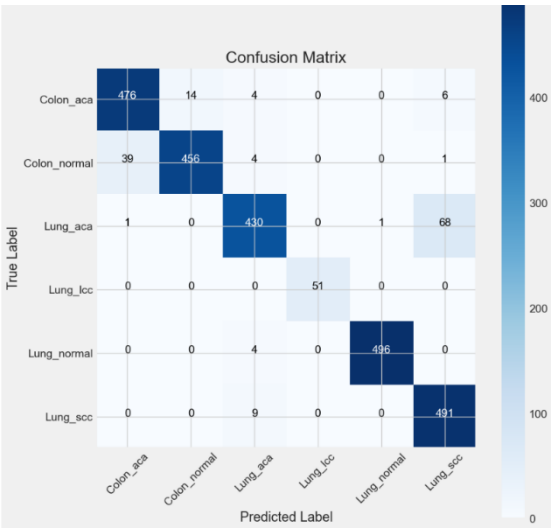

Fig-12: Confusion Matrix of Different Image Categories for Validation Images

	precision	recall	f1-score	support
Colon_aca	0.92	0.95	0.94	500
Colon_normal	0.97	0.91	0.94	500
Lung_aca	0.95	0.86	0.90	500
Lung_lcc	1.00	1.00	1.00	51
Lung_normal	1.00	0.99	0.99	500
Lung_scc	0.87	0.98	0.92	500
accuracy			0.94	2551
macro avg	0.95	0.95	0.95	2551
weighted avg	0.94	0.94	0.94	2551

Table-4: Accuracy for Sigmoid Activation function and RMSprop Optimizer.

V. CODE

All the codes are uploaded in this repository: https://github.com/RezuanIslam/Lung-and-Colon-Cancer-Classification-and-Detection-using-Convolutional-Neural-Network.git

VI. DISCUSSION

At first, we used different activation functions and different optimizers to evaluate the performance of our own model with the help of MobileNet to demonstrate the result in the used dataset. The best result we got from using the sigmoid activation function and RMSprop optimizer in the last layer where we got 95.34% training accuracy and 93.67% validation accuracy. Using the softmax activation function and RMSprop optimizer we got the lowest accuracy which is 77.0% training accuracy and 76.52% validation accuracy. When the softmax activation function and Adam optimizer were used the model got 94.11% training accuracy and 93.31% validation accuracy. Which is the second-best result among these three.

The confusion matrix shown in three figures for different models depicts the true label vs. the predicted label of the images for the validation data in given labeled categories. Tables 2, 3, and 4 show the precision, recall, and f-score for the different histopathology image categories. The formula to calculate the given metrics is explained in section III.

VII. CONCLUSION

This research work presents lung and colon cancer classification and detection using histopathological images. A convolutional neural network was implemented to classify six different categories of cancer large cell carcinoma, adenocarcinoma, squamous cell carcinoma, normal tissue, colon adenocarcinoma, and colon benign tissue using the MobileNet model. The best result our model was able to achieve 95.34% and 93.67% of training and validation accuracy. The precision, recall, and f1-socre were calculated, and a confusion matrix plot was drawn to measure the model performance.

There was some limitation of our work. The category of the large cell carcinoma images dataset was not big enough. We can apply our model on a bigger dataset in the future. The network was not big enough to detect complex features. We can make our network more complex and bigger which can detect more complex features and extract them for more accuracy. Although more accuracy might be gained using other activation functions and optimizers by using our model. But it is still great to get an accuracy of 95.34%.

VIII. REFERNCES

[1] Aya Hage Chehade, Abdallah, N., Marion, J.-M., Mohamad Oueidat, & Chauvet, P. (2022). Lung and colon cancer classification using medical imaging: a feature engineering approach. 45(3), 729–746. https://doi.org/10.1007/s13246-022-01139-x

[2] Bray F., Ferlay J., Soerjomataram I., Siegel R.L., Torre L.A., Jemal A., Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA: a cancer journal for clinicians, 2018. 68, 6.394–424.

[3] Bray F., Ferlay J., Soerjomataram I., Siegel R.L., Torre L.A., Jemal A., Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA: a cancer journal for clinicians, 2018. 68, 6.394–424.

[4] Mesut Togaçar, Disease type detection in lung and colon cancer images using the complement approach of inefficient sets, Computers in Biology and Medicine, Volume 137, 2021, 104827, ISSN 0010-4825, https://doi.org/10.1016/j.compbiomed.2021.104827.

[5] Sánchez-Peralta L.F., Bote-Curiel L., Picón A., Sánchez-Margallo F.M., Pagador J.B., Deep learning to find colorectal polyps in colonoscopy: A systematic literature review, Artificial Intelligence in Medicine, Volume 108, 2020, 101923, ISSN 0933-3657, https://doi.org/10.1016/j.artmed.2020.101923.

[6] K. Yu, C. Zhang, G. Berry, et al. "Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features." Nat Commun 7, 12474 (2016), doi: 10.1038/ncomms12474

[7] L. G. Collins., C. Haines, R. Perkel & R. E. Enck. "Lung cancer: diagnosis and management." American family physician vol. 75, 1 (2007): 56-63

[8] Bukhari S.U.K., Asmara S., Bokhari S.K.A., Hussain S.S., Armaghan S.U., SHAH S.S.H., The Histological Diagnosis of Colonic Adenocarcinoma by Applying Partial Self Supervised Learning, medRxiv 2020, https://doi.org/10.1101/2020.08.15.20175760.

[9] Hatuwal B.K., Thapa H.C., Lung Cancer Detection Using Convolutional Neural Network on Histopathological Images, International Journal of Computer Trends and Technology, Volume 68 Issue 10, 21-24, October 2020, doi: 10.14445/22312803/IJCTT-V68I10P104

[10] S. Mangal, A. Chaurasia, A. Khajanchi, Convolution neural networks for diagnosing colon and lung cancer histopathological images, arXiv 2020, arXiv:2009.03878.

[11] B. K. Hatuwal and H. C. Thapa, “Lung cancer detection using convolutional neural network on histopathological images,” International Journal of Computer Trends and Technology, vol. 68, no. 10, pp. 21–24, Oct. 2020, doi: 10.14445/22312803/ijctt-v68i10p104.

[12] M. Nishio, K. Fujimoto, H. Matsuo, C. Muramatsu, R. Sakamoto, and H. Fujita, “Lung cancer Segmentation with transfer learning: usefulness of a pretrained model constructed from an artificial dataset generated using a generative adversarial network,” Frontiers in Artificial Intelligence, vol. 4, Jul. 2021, doi: 10.3389/frai.2021.694815.

[13] O. Ronneberger, P. Fischer, and T. Brox, “U-NET: Convolutional Networks for Biomedical Image Segmentation,” in Lecture Notes in Computer Science, 2015, pp. 234–241. doi: 10.1007/978-3-319-24574-4_28.

[14] M. Beldar, S. Rajguru, S. Suman, T. B. Patil “A Comparative Analysis For Early Diagnosis of Lung Cancer Detection and Classification By Ct Images Processing Using Resnet-50 Model of Cnn”, 2023

[15] M. Schultheiss et al., “A robust convolutional neural network for lung nodule detection in the presence of foreign bodies,” Scientific Reports, vol. 10, no. 1, Jul. 2020, doi: 10.1038/s41598-020-69789-z.

[16] Borkowski AA, Bui MM, Thomas LB, Wilson CP, DeLand LA, Mastorides SM. Lung and Colon Cancer Histopathological Image Dataset (LC25000). arXiv:1912.12142v1 [eess.IV], 2019

[17] MOHAMED HANY. Chest CT-Scan images Dataset, 2020

- **CONTRIBUTION**

	REZUAN ISLAM	MD. MASHRURUL HAQUE	TAMANNA AKTER	DAVASIS SARKAR	Contribution (%)
	XX-XXXXX-X	*XX-XXXXX-X*	*XX-XXXXX-X*	*XX-XXXXX-X*	
Conceptualization	25%	25%	25%	25%	100 %
Data curation	50%	50%	0%	0%	100 %
Formal analysis	25%	25%	25%	25%	100 %
Investigation	25%	25%	25%	25%	100 %
Methodology	60%	40%	0%	0%	100 %
Implementation	30%	70%	0%	0%	100 %
Validation	50%	50%	0%	0%	100 %
Theoretical derivations	0%	0%	50%	50%	100 %
Preparation of figures	80%	20%	0%	0%	100 %
Writing – original draft	25%	25%	25%	25%	100 %
Writing – review & editing	40%	20%	20%	20%	100 %