“LUNG CANCER DETECTION” USING CONVOLUTIONAL NEURAL NETWORK (CNN)

Sonu Kumar[1], Manisha Minz[2] and Sushant Lawrence Guria[3]

sonu82256@gmail.com[1], minzmanisha93@gmail.com[2] and guriasushant@gmail.com[3]

DEPARTMENT OF MBC, MAULANA AZAD NATIONAL INSTITUTE OF TECHNOLOGY, BHOPAL, INDIA

Submitted to :- DR. G.S. THAKUR MCA523.manit+GST@gmail.com
DEPARTMENT OF MBC, MAULANA AZAD NATIONAL INSTITUTE OF TECHNOLOGY, BHOPAL, INDIA

Abstract *— Lung Cancer is one of the leading life taking cancer worldwide. Early detection and treatment are crucial for patient recovery. Medical professionals use histopathological images of biopsied tissue from potentially infected areas of lungs for diagnosis. Most of the time, the diagnosis regarding the types of lung cancer are error-prone and time-consuming. Convolutional Neural networks can identify and classify lung cancer types with greater accuracy in a shorter period, which is crucial for determining patients' right treatment procedure and their survival rate. Benign tissue, Adenocarcinoma, and squamous cell carcinoma are considered in this research work. The CNN model training and validation accuracy of 99.04 and 92.33 percentage are obtained.*

Keywords — *Convolutional Neural Network (CNN), Machine Learning, Lung Cancer, Histopathological Image*

I. INTRODUCTION

The abnormal growth of cells in human Lung is called as Lung Cancer. Lung cancer is one of the most serious diseases in the world today. Lung cancer is prominent cancer among both men and women, making up almost 25% of all cancer deaths (i.e; breast, prostate, and colon cancer put together). The primary cause of death from lung cancer, about 80% is from smoking. Lung cancer in non-smokers can be caused by exposure to radon, second-hand smoke, air pollution, or other factors like workplace exposures to asbestos, diesel exhaust, or certain other chemicals. In addition, genetic factors also have a major contribution to lung cancer. Uncontrolled magnification of tissue creates lung cancer.

Lung cancer can be cancerous or noncancerous. Lower-grade cancers are classified as Grades I and II. In some cases, cancer grades III and IV are regarded to be of higher severity. Normal cells are small and confined, whereas cancer-affected cells are rapidly forming and can be easily spotted. These cells appear to be aberrant and dissimilar to regular cells. This type of cell grows quickly and is more prone to spread.

Various tests like imaging sets (x-ray, CT scan), Sputum cytology, and tissue sampling (biopsy) are carried out to look for cancerous cells and rule out other possible conditions. While performing the biopsy, evaluation of the microscopic histopathology slides by experienced pathologists is indispensable to establishing the diagnosis and defines the types and subtypes of lung cancers. For pathologists and other medical professionals diagnosing

lung cancer and the types is a time consuming process. There is a significant change the cancer types are misdiagnosed, which directs to incorrect treatment and may cost patients' lives.

Machine Learning (ML) is a subfield of Artificial Intelligence (AI) that allows machines to learn without explicit programming by exposing them to sets of data allowing them to learn a specific task through experience . In previous research papers, most of the authors considered using x-rays, CT scans images with machine learning techniques such as Support Vector Machine (SVM), Random Forest (RF), Bayesian Networks (BN), and Convolutional Neural Network (CNN) for lung cancer detection and recognition purpose. Some papers also considered using histopathological images, but they distinguish between carcinomas and non-carcinomas images and with lower accuracy. This research paper has considered using Convolutional Neural Network (CNN) architecture to classify the normal, Adenocarcinoma, and squamous cell carcinomas. We have not found other papers using the CNN model to classify only the given three different histopathological images and the given model's accuracy.

The methodology and settings used are described briefly in Section II & III. Similarly, the research's obtained output is explained and shown with plots and tables in Section IV. The conclusion of the paper is explained in Section V and cited sources mentioned in the References section.

II. About Dataset (Materials)

This dataset contains 15,000 histopathological images with 3 classes. All images are 768 x 768 pixels in size and are in jpeg file format.

The images were generated from an original sample of HIPAA compliant and validated sources, consisting of 15,000 total images of lung tissue (5,000 benign lung tissue, 5,000 lung adenocarcinomas, and 5,000 lung squamous cell carcinomas) There are five classes in the dataset, each with 5,000 images, being:

- Lung benign tissue
- Lung adenocarcinoma
- Lung squamous cell carcinoma

III. METHODOLOGY

Our proposed system followed data acquisitions, data formatting, model training, testing, and prediction, described in the below sections.

A. Data Acquisition

The histopathology images are obtained from LC15000 Lung and colon histopathological image dataset. Three classes of benign tissue, Adenocarcinoma, and squamous carcinoma cells of lungs with 5000 histopathology images in each category, are considered for our work.

B. Data Formatting

The obtained dataset was RGB color histopathology images with .jpeg format. The images were resized to maintain a uniform aspect ratio of one with (256, 256) pixel size for the CNN operation. We have implemented the image acquisition technique like horizontal and vertical flip and zooming to increase the image number and variation in the data pattern. The neural network tends to over-fit in case of a limited number of training data samples trained with a higher number of

epochs. Fig. 1(a) and Fig. 1(b) show Adenocarcinoma's histopathology image and its augmented images, respectively, with the horizontal and vertical flip and a zoom range of 0.2 applied.

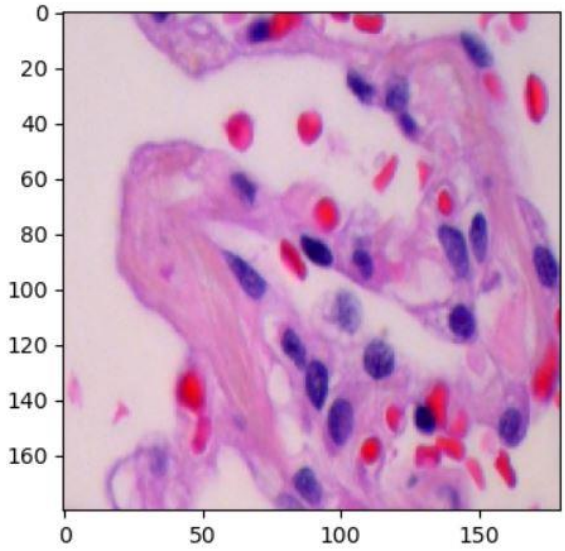

Fig 1(a): Histopathology Image of Adenocarcinoma

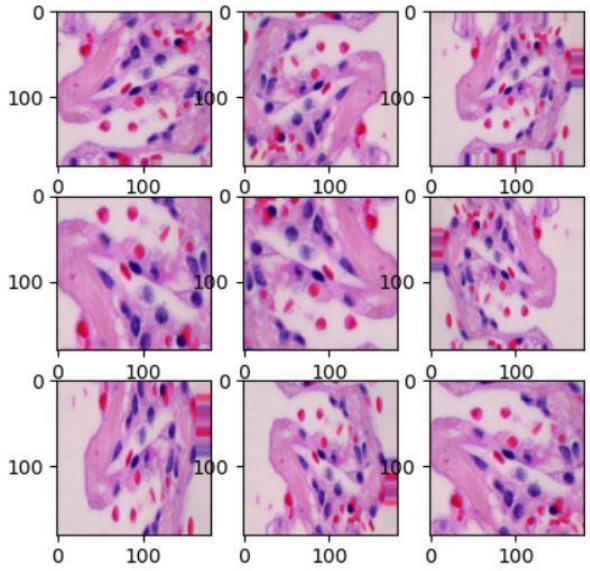

Fig 1(b): Corresponding Augmented Histopathology Images of Adenocarcinoma

C. Model Training, Testing, and Prediction

A liner stack of layers was used to create the Convolutional Neural Network (CNNs or ConvNets) for the image classification and recognition. Training and testing images were passed through convolutional layers with kernel filters, max pooling, and fully connected layers. The softmax function was applied to classify the given object.

A neural network with three hidden layers, one input layer, and one fully connected layer was implemented for this task. Images are split in a ratio of 80:20 for training and validation purposes. Images of (256, 256) pixel size were passed to the input layer. Kernel matrix of (3, 3) with (ReLU(x) = max (0, x)) as an activation function was applied in each convolutional layer. Max pooling size of (2, 2) was implemented to reduce the computation parameters in the next convolution layer. A dropout value of 0.1 was applied to the model. A dense value of three with the sigmoid activation function was used to obtain the class probabilities for final output classes. An adaptive moment estimation (Adam) optimizer was used to calculate the learning rates for different parameters. Loss function calculates the discrepancy between the predicted output and the labeled output for the given input; categorical cross-entropy (CE) was used as a loss function for this task, which is calculated as:

$$CE = -\log\left(\frac{e^{S_p}}{\sum_j^C e^{S_j}}\right)$$

C is the number of output class, S_p is the CNN score of the given positive class, and S_j is the score inferred by the net for each class C.

The performance of the developed CNN model was measured using the confusion matrix plot, and the metrics accuracy,

precision, recall, and f1-score were also calculated as below:

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

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

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

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

Where TP, FP, FN, and TN represents the output measures as true positive, false positive, false negative, and true negative values for the training and validation images of the models.

IV. RESULT AND DISCUSSION

The images were trained for 10 epochs with batch size 64. The model achieved a training accuracy of 99.04% and a validation accuracy of 92.33% in the final epoch.

Below, Fig. 2(a) and Fig. 2(b) shows the plot of model accuracy vs. epoch and model loss vs. epoch for training and validation images.

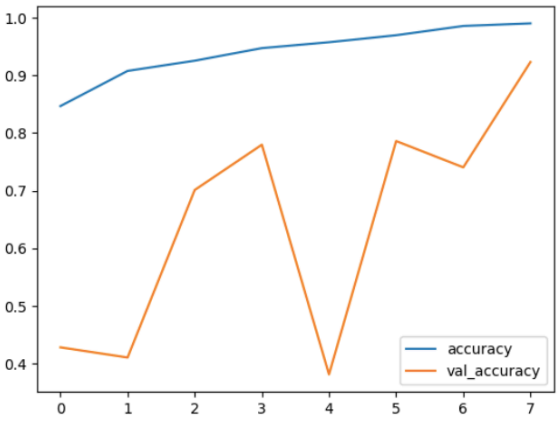

Fig 2(a): Plot of Model Training and Validation Images

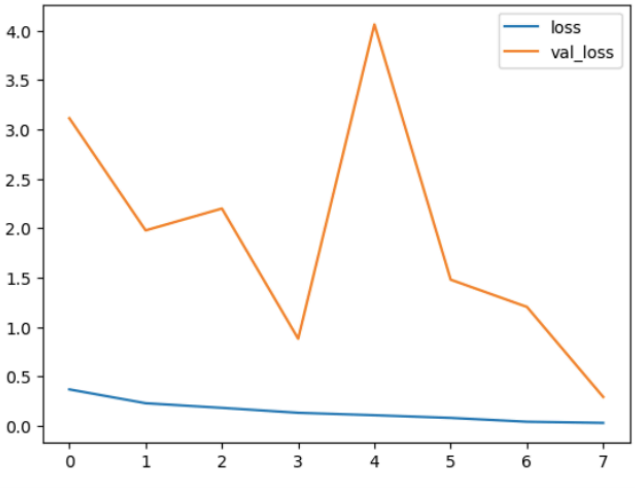

Fig 2(b): Plot of Model Loss vs. Epoch for Training and Validation Images.

Table I

Precision, Recall, and F1-Score of Model for Different Categories

Category	Precision	Recall	F1-score
lung_aca	0.85	0.94	0.89
lung_n	1.00	0.93	0.97
lung_scc	0.94	0.90	0.92

The table shows the precision, recall, and f-score for the different histopathology image categories. The formula to calculate the given metrics is explained show in Section III-C.

```
array([[927,   0,  60],
       [ 65, 912,   0],
       [105,   0, 931]])
```

Fig (3): Fig(3). Confusion Matrix of Different Image Categories for Validation Images

The confusion matrix shown in Fig. 3 depicts the true label vs. the predicted label of the images for the validation data in given labeled categories.

V. CONCLUSION

This research work presents lung cancer detection using histopathological images. A convolutional neural network (CNN) was implemented to classify an image of three different categories benign, Adenocarcinoma, and squamous cell carcinoma. The model was able to achieve 99.04% and 92.33% of training and validation accuracy. The precision, f1-socre, recall were calculated, and a confusion matrix plot was drawn to measure the model performance.

VI. REFRENCES

1. 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

2. Vikas Pal Veelon and G.S.Thakur. “K-NN Pridict” Tool to pridict secondary structure of Proteins. DEPARTMENT OF MBC,MAULANA AZAD NATIONAL INSTITUTE OF TECHNOLOGY, BHOPAL, INDIA

3. R. Pandian, V. Vedanarayanan, D.N.S. Ravi Kumar, R. Rajakumar, Detection and classification of lung cancer using CNN and Google net

4. Lung Cancer Detection Using Convolutional Neural Network, Dunke Siddhi1, Tarade Swapna2, Waghule Pratiksha3, 1Dunke siddhi, JCOE kuran, 2Tarade Swapna, JCOE kuran, 3Waghule Pratiksha, JCOE kuran, 4Prof. Kolase Sachin, Department of Computer Engineering, JCOE, Kuran, Maharashtra.

5. Lung Cancer Detection Using Convolutional Neural Network on Histopathological Images Bijaya Kumar Hatuwal1, Himal Chand Thapa2 1Himalaya College of Engineering (of Tribhuvan University), Chyasal, Lalitpur, Nepal 2Lecturer, Himalaya College of Engineering (of Tribhuvan University), Chyasal, Lalitpur, Nepal.

6. Data Mining: Concepts and Techniques 3rd Edition - June 9, 2011, Authors: Jiawei Han, Micheline Kamber, Jian Pei

7. Data Science from Scratch, 2nd Edition by Joel Grus, Released April 2019, Publisher(s): O'Reilly Media, Inc.