International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 01 | Jan 2022 www.irjet.net p-ISSN: 2395-0072

Preliminary Lung Cancer Detection using Deep Neural Networks

Neha Mayekar[1], Shreya Pattewar[2], Shubham Patil[3], Ajay Dhruv[4]

[1]*Student, Information Technology, Vidyalankar Institute of Technology, Maharashtra, India*
[2]*Student, Information Technology, Vidyalankar Institute of Technology, Maharashtra, India*
[3]*Student, Information Technology, Vidyalankar Institute of Technology, Maharashtra, India*
[4]*Prof., Dept. of Information Technology, Vidyalankar Institute of Technology, Maharashtra, India*

Abstract - *In many countries around the world, lung cancer is the leading cause among cancer deaths. Tumors can be benign or malignant. "Cancer" here refers to malignant tumors. Causes of lung cancer include passive smoking, direct smoking, exposure to certain toxins and family history. In this paper, Convolutional Neural Network is proposed for classification of histopathological lung cancer tissue images; also, performance of proposed CNN model is compared with other pre trained architectures. Successful detection or prediction of lung cancer in early-stage is very important. It will effectively reduce treatment costs which will be incurred in the future. It can reduce or eliminate risk of surgery and even increase survival rate. Therefore, the lung cancer detection systems are effective methods for early-stage detection of lung cancer, besides being easy, cost effective and time saving solutions. Hence, the system plays a vital role in the diagnosis process.*

Key Words: **Deep learning, lung cancer, CNN, VGG16, Inception v3.**

1.INTRODUCTION

Lung cancer is a common type of cancer occurring in the lung. It causes damage to the lung tissues. A person is said to have lung cancer when there is irregular growth of cells in the lung. It may begin in one or both lungs. Coughing up blood, shortness of breath, cough are some of the common lung cancer symptoms. It most often occurs in the people who smoke. Prime cause of lung cancer is cigarette smoking. Major cause of deaths occurring in the United States is cigarette smoking [14, 15, 16]. As per the estimation provided by the World Health Organization in May 2020, more than 8 million people die due to tobacco use annually. Among these, approximately seven million die due to direct tobacco use, while 1.2 million people die because of being exposed to second-hand smoke [17].

According to [18], cigarette smoke consists of more than 4000 chemicals. Majority of these chemicals are cancer causing. A person has a 20-25 times greater risk of developing lung cancer who smokes more than one pack of cigarettes per day than someone who has never smoked and can increase the chance of survival by 73% due to early detection [19]. Tobacco use or smoking causes about 90% of lung cancer related deaths [1]. Though smoking remains the prime factor of causing lung cancer there are other factors as well. Environmental pollution (mainly air), excess alcohol consumption, exposure to tobacco smoke (passive smoking), among others also contribute to development of lung cancer.

If lung cancer is predicted or detected early, then it will help a lot of people. This could work as an effective preventive strategy. The techniques used in the diagnosis lung cancer are Computed Tomography (CT), Chest Radiography (x-ray), Sputum Cytology, Magnetic Resonance Imaging (MRI scan) etc.

However, most of these techniques are costly and time consuming. Besides the chance of survival of the patient is greatly affected due to late or delayed cancer detection. [21] In order to overcome this there is a need of a technical solution for early diagnosis, which will provide faster results. It will greatly assist the medical staff [2].

In this system, deep learning techniques are used for early detection of cancer. In the system whole slide images are processed using the Convolutional Neural Network algorithm for classification of cancer and non- cancer images [3]. [23] In this early detection system, feature extraction is handled by the algorithm itself. Using these features, the system will predict lung cancer. The motivation of designing the system is to provide a fast and cost-effective solution for early detection of lung cancer. The detection will be based on some factors and thresholds which are explained later [22]. Deep Learning is prominently used in the field [12, 13] of medical image Processing.

Section 2 discusses literature survey. Section 3 provides information on the dataset used in this paper. Section 4 discusses the proposed system. Section 5 discusses conclusion and future scope.

2. LITERATURE SURVEY

Convolutional neural networks are widely used in Image Classification Systems [4]. Most of the time features are extracted from the top layer of CNN Model and then they are further utilized for classification. One problem with this technique is that it does not always provide relevant information. This paper makes use of a pre trained CNN model for feature extraction. They proposed a system in which feature extraction is done with the help of the second and third layer of the already learned CNN Model and the extracted features are further provided to another CNN Model. The Proposed method improves the result of existing models.

Nowadays, medical imaging is widely being used in histopathological lung cancer classification [5]. The paper presents a very effective CNN based architecture for histopathological lung cancer image classification. The paper presents a CNN architecture which has the following structure: one input layer, multiple blocks of convolutional layers, drop-out layers and max-pooling layer combination and fully connected layer. In total there are 16 convolutional layers, 05 dropout layers, 05 max-pooling layers and 01 fully connected layer. Therefore, the proposed architecture confirms the higher performance against existing models.

In [6], the dataset of MRI images is used. However, the dataset used is small in size, hence the method of transfer learning is adapted. Transfer learning indicates use of deep pre-trained convolutional Neural Network architectures rather that building the entire model based on trial and error. The paper uses nine pre-trained architectures (AlexNet, GoogleNet, VGG-19, VGG16, ResNet-18, ResNet-50, ResNet-101, ResNet-inception-v2, SENet), on the dataset. The results are generated over 25 epochs, 50 epochs and 90 epochs separately. A comparative study is then performed over the results. In conclusion, three epochs are chosen and these show that VGG16, VGG19 and some layers of AlexNet perform better than ResNet and ResNet-Inception-v2. They achieved 98.71%, 98.55% and 98.55%, accuracy respectively. 98.55% accuracy was achieved by both AlexNet and VGG16 with 50 epochs. But VGG16 consumed more time compared to AlexNet.

In [7], the evolution of the neural networks over the period of time is explained. The paper provides an overview of how deep learning can be used in medical image analysis by explaining the techniques of Classification, detection, and segmentation on pulmonary medical images. The process of detecting the pulmonary nodule using deep learning is also explained. These methods are implemented on various lung diseases such as pulmonary nodule diseases, pulmonary embolism, pneumonia and interstitial lung disease and an overview is presented.

In [8], chest X-ray dataset is used in order to classify pneumonia. They have applied three techniques on this dataset, first is linear support vector machine model with local rotation and orientation free features. Second is transfer learning with the help of two convolutional neural network models. Third technique used is the capsule network. Since the used dataset was not big enough, various kinds of augmentation are applied on the same data in order to improve performance. From these three experiments they concluded that CNN based transfer learning gives best results out of three. Even capsule networks perform better than ORB and SVM classifiers.

In [9], X-ray images dataset is used. The dataset consists of various categories of images such as bacterial pneumonia, positive Covid-19 disease, and normal incidents. The aim of the paper is to automate detection of coronavirus disease. Specifically, they have adopted a transfer learning approach to automate the process. They have used 6 pre-trained architectures VGG19, MobileNet v2, Inception, Xception, Inception Resnet v2. They concluded that VGG19 and MobileNet v2 gives best results for their desired target. They have also compared their results on various parameters as accuracy, sensitivity, and specificity.

3. DATASET

This paper uses dataset named "Histopathological Cancer Detection" which has been taken from Kaggle. Dataset consists of whole slide images of cancerous and non-cancerous tissue [10]. Dataset has 2,20,025 RGB images for training with respective labels and another 57,458 imagesfor testing. Out of 2,20,025 training images, 41% are cancerous images and remaining of non-cancer.

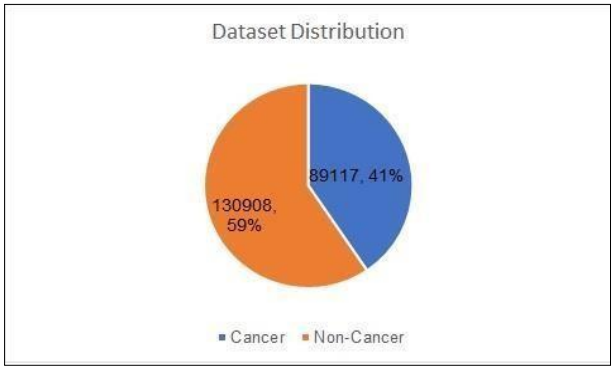

Fig -1: Dataset Distribution

From the pie chart shown in fig. 1**,** it can be interpreted that the dataset is imbalanced. Since an imbalanced dataset affects accuracy, hence it is necessary to balance the dataset. To balance out the dataset, the concept of down sampling(under sampling) is used by reducing the size of non-cancer class which is in abundance. Hence 89000 images from both classes are used, therefore total 178000 images have been used, 10% out of this, that is 17800 images are used for validation and the other 90% used for training models.

Table -1: Dataset Distribution

Type / Division	Cancer	Non-Cancer	Total
Training	80100	80100	160200
Validation	8900	8900	17800
Total			178000

4. PROPOSED SYSTEM

The convolutional neural network (CNN) is a specialized type of neural network model. This architecture is designed for working with one, two and three-dimensional image data [11, 20]. CNN is a sequence of multiple layers. These layers can be

convolution layer, non-linear activation layer, dropout layer, pooling layer, fully connected layer.

Convolution Layer:

The prime component of the network is the convolutional layer; hence, the architecture is named after it. This layer is responsible for performing linear operation known as convolution. Convolution is nothing but multiplying the input with a set of weights. In the proposed architecture convolution operation multiplies two-dimensional array of weights (filters or kernels) with the input data.

Activation Layer:

Convolutional neural networks are designed to extract and analyze features from images. Basic features are extracted at initial level, however very high-level features are extracted at deeper levels. The purpose of introducing this layer is to introduce nonlinearity, as nonlinear functions are good at generating generalized features and makes training faster. Most used activation layer is ReLU (Rectified Linear Unit). ReLU returns 0 if input value is 0 or less than 0 else it returns input value as it is as output. Other activation functions are tanh, sigmoid, and SoftMax.

Pooling Layer:

Most relevant features can be extracted using pooling layers. Pooling layers also helps in reducing complexity of the network by performing spatial dimensionality reduction. Average pooling and max pooling are the most used pooling layers. Max pooling layer selects maximum value from region of input where kernel/filter is placed, however average pooling layer takes average of all values from region of input where kernel/filter is placed.

Dropout layer:

This layer is used to prevent the networks from overfitting by randomly deactivating some neurons while training.

If we compare CNN with other classification algorithms, CNN requires relatively less pre-processing. In traditional algorithms filters are hand-engineered. CNN has the ability to eliminate this process and learn characteristics/features from the input data itself. In the last few years, CNN has proved itself to be the dominant method due to its improved accuracy in fields such as computer vision.

CNN works best for classification of image dataset. However, CNN can be used in two ways, which are transfer learning and the other one is using a specific convolutional network.

There are various pre-trained architectures which uses CNN based transfer learning such as VGG16, Inception V3 [24], Resnet Inception-ResNet, Xception, and so on. All of these architectures are trained on ImageNet dataset. Feature extraction and fine tuning are the two ways to use pre-trained architectures. In this paper, for feature extraction VGG16 and Inception V3 are used. Later these extracted features are given as input to neural networks and random forest algorithms which work as classifiers. The performance of these models are compared with proposed CNN architecture and are shown in fig. 2

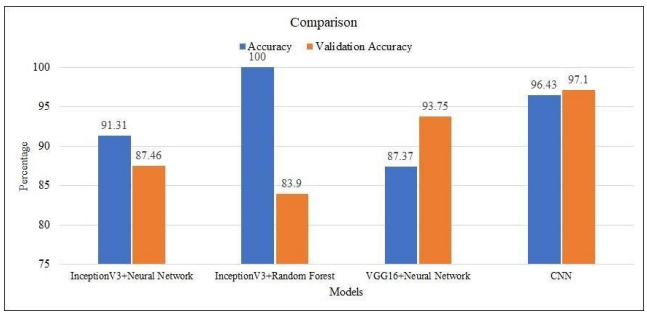

Fig -2: Comparison Chart

The table 2, shows that either pre- trained architecture overfits or underfits the given dataset. Overfitting refers to a model that performs well on the training data, but performance is degraded on validation data. Underfitting refers to a model that can neither model the training datanor generalize to new data. To put it simply, overfitting occurs when the validation metrics are far worse than the training metrics. On the contrary, underfitting occurs when training metrics are far worse than validation metrics.

Table -2: Overview of different models Proposed CNN

Models	Training Accuracy	Validation Accuracy	Comparison	Outcome
InceptionV3 + Neural Network	91.31	87.46	Training accuracy > Validation accuracy	Overfitting
InceptionV3 + Random Forest	100	83.9	Training accuracy >> Validation accuracy	Overfitting
VGG16 + Neural Network	87.37	93.75	Training accuracy < Validation accuracy	Underfitting
CNN	96.43	97.1	Training accuracy almost equal Validation Accuracy	Fits well

The proposed architecture is a Convolutional Neural Network classification model. It performs binary classification on histopathological lung cancer images. The proposed architecture consists of the following elements: three blocks of convolutional layer, one max pooling layer and one dropout layer. Each element is used multiple times to build the complete architecture. Hence, the final model has twelve convolutional layers, four dropout layers, three max pooling layers, one flatten layer and finally one dense layer. The first block consists of four convolutional layers, which in turn has thirty-two filters, one max pooling layer and a drop-out layer. Each filter is of size 3 x 3 with ReLU activation function and valid padding. The max pooling layer is of size 2. The drop-out layer has a significant probability of 0.3.

The output from the first block is passed to the next block. The next block has the following elements: convolutional

layers which has sixty-four filters each of size 3x3 followed by one max pooling layer with filter size of 2x2 and finally a drop-out layer with probability 0.3. The same structure is repeated in the last block with the only difference of 128 filters in the convolutional layer. The output from this layer is passed on to the flatten layer followed by dense layer and finally a drop-out with 0.3 probability. The dense layer has 256 units and ReLu activation function. ReLu activation function is a nonlinear function. It does not activate all layers at the same time; hence it generalizes well. At the last dense layer is connected with 2 nodes and a SoftMax activation function. Since this paper focuses on binary classification, hence 2 nodes are used in the output layer.

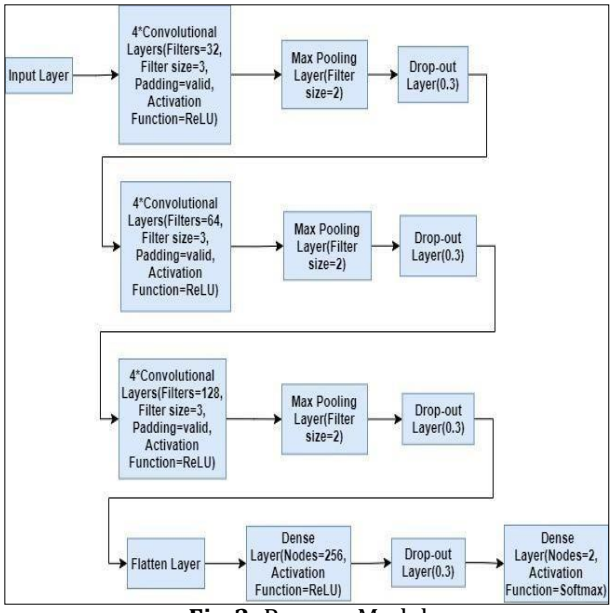

Fig-3: Process Model

When the above CNN model is trained on histopathological lung cancer images for 30 epochs, the model training history obtained are shown as follows:

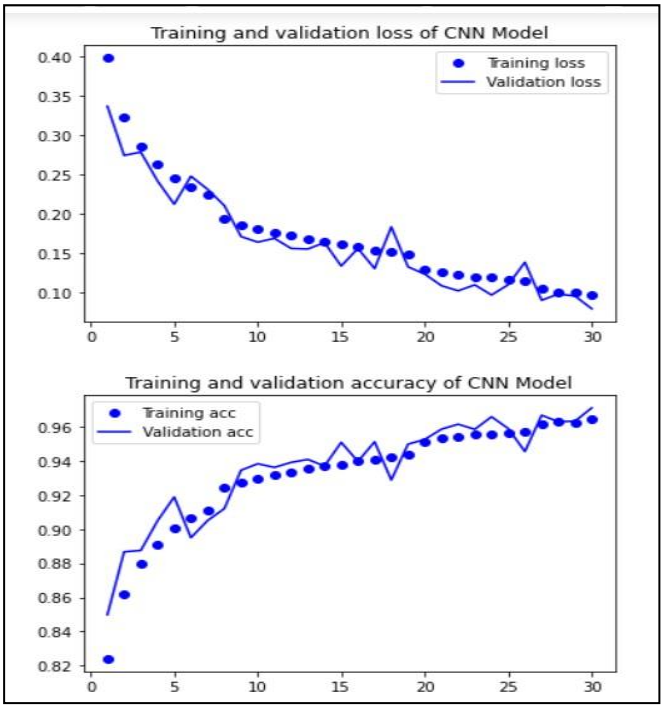

Fig -4: Model Training History

From the above plots, it is seen that the model has comparable performance on both train and validation datasets.

Table -3: Results

Loss	9.70%
Accuracy	96.43%
Validation Loss	8.00%
Validation Accuracy	97.10%

From the table no 3, it can be observed that accuracy and validation accuracy are quite good. Also, both accuracies are very close, which proves that the proposed CNN model very well generalizes the used dataset.

Performance of the proposed CNN model is validated by generating a confusion matrix which is shown in fig. 5.

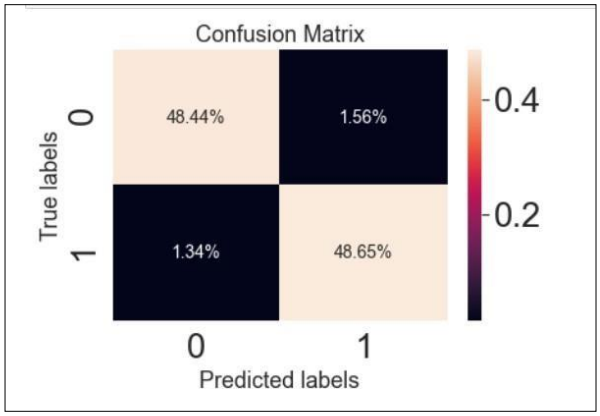

Fig -5: Confusion Matrix

Generated confusion matrix is of size 2*2 because there are only two classes of images in the dataset used which are cancerous and non-cancerous. In this matrix, predicted labels are shown on x-axis, and true labels are depicted on y-axis. Furthermore, to judge the performance of the proposed model, different performance measures like accuracy, precision, specificity, f1 score are calculated and shown in table no 4.

Table -4: Performance measure of proposed CNN model

Performance measure	Result
Accuracy	97.09%
Precision	96.89%
Recall	97.31%
F1 Score	97.09%
Specificity	96.88%

Ideally a good classifier should have all measures mentioned in table 4 should be 100%. From the table, it can be observed that all the measure values are greater than 96%. This proves that the proposed system classifies the given dataset very well.

5. CONCLUSION

The dataset used contains RGB images with two classes. In this paper, three CNN based architectures are used which are proposed CNN model, and two pre-trained architectures which are VGG16 and Inception V3. After comparing these, it has been observed that the proposed CNN model gives best performance than pre trained architectures which make use of CNN based transfer learning. The performance of proposed architecture has been analyzed by generating confusion matrices and by computing different performance measures which are accuracy, precision, recall, F1- score and specifically, all of these give acceptable results. As a future extension to this project other pre- trained architectures like Resnet50, Xception, VGG19, MobileNet, etc. can also be tested on the dataset. Further, the cancerous patch can be marked on the histopathological slide image of lung tissue which will help locate the exact areas

REFERENCES

[1] Neha Panpaliya, Neha Tadas, Surabhi Bobade, Rewati Aglawe, Akshay Gudadhe, "A SURVEY ON EARLY DETECTION AND PREDICTION OF LUNG CANCER", International Journal of Computer Science and Mobile Computing Vo. 4 Issue, January- 2015

[2] J. Liu, H. Chang, C. Wu, W. S. Lim, H. Wang and J. R. Jang, "Machine Learning Based Early Detection System of Cardiac Arrest", 2019 International Conference on Technologies and Applications of Artificial Intelligence(TAAI), Koahsiung, Taiwan, 2019

[3] Umesh D R. Dr. B Ramachandra, "Association Rule Mining Based Predicting Breast Cancer Recurrence on SEER Breast Cancer Data", International Conference on Emerging Research in Electronics, 2015

[4] Junho Yim, Jeongwoo Ju, Heechul Jung and Junmo Kim, "Image Classification Using Convolutional Neural Networks With Multi-stage Feature", Springer International Publishing Switzerland 2015

[5] Venubabu Rachapudi, G. Lavanya Devi, "Improved convolutional neural network based histopathological image classification", part of Springer Nature 2020

[6] Chelghoum R., Ikhlef A., Hameurlaine A. Jacquir S. (2020) Transfer learning Using Convolutional Neural Network Architectures for Brain Tumor Classification from MRI Images, vol 583. Springer, Cham.

[7] Jiechao Ma, Yang Song, Xi Tian, Yiting Hua, Rongguo Zhang, Jianlin Wu, "Survey on deep learning for pulmonary medical".

[8] S. Yadav, M. Jadhav, Deep Convolutional neural network based medical image classification for disease diagnosis, 2019.

[9] Ioannis D. Apostolopoulos, Tzani A. Mpesiana, Covid-19 Automatic detection from X-ray images utilizing transfer learning with convolutional neural networks, 2020

[10] M. Šarić, M. Russo, M. Stella and M. Sikora, "CNN- based Method for Lung Cancer Detection in Whole Slide Hitopathological Images", 2019 4th International Conference on Smart and Sustainable Technologies, Split, Croatia, 2019.

[11] Rikiva Yamashita, Mizuho Nishio, Richard Kinh Gian Do, Kaori Togashi, "Convolutional neural networks: an overview and application in radiology", Insights into Imaging - 2018

[12] Dinggang Shen, Guorong Wu, and Heung-Il Suk, "Deep learning in medical analysis", 2017

[13] T. Turki, "An empirical study of machine learning algorithms for cancer identification," 2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC), Zhuhai, 2018, pp. 1-5.

[14] Report on Adult Cigarette Smoking in the United States, 2018.

[15] Report on the global tobacco epidemic 2011 by WHO.

[16] Prajakta Belsare1, Volkan Y Senyurek1 , Masudul H Imtiaz1 , Stephen Tiffany2 , and Edward Sazonov1*, Senior Member, IEEE "Computation of Cigarette Smoke Exposure Metrics from Breathing", IEEE, 2020

[17] Report on Tobacco 2020 by WHO

[18] "Health Effects of Cigarette Smoking".

[19] Report by University of Liverpool, 2016

[20] Andrew Zisserman, Karen Simonyan, Very Deep convolutional networks for large scale Image Recognition.

[21] Rebecca Siegel, Deepa Naishadham and Ahmedin Jernal Cancer statistics, 2013

[22] Pal R, Saraswat M, Histopathological, Image classification using enhanced bag-of-feature with spiral biogeography-based optimization.

[23] Pal R, Saraswat M, Enhanced bag of features using Alexnet and improved biogeography based optimization for histopathological image analysis. Eleventh international conference on contemporary computing.

[24] Deep Learning with Python (1st. ed.) by Francois Chollet.