International Journal for Research in Engineering Application & Management (IJREAM)
ISSN : 2454-9150 Vol-05, Issue-03, June 2019

Lung Cancer Detection using Transfer Learning

[1]Mr. Chaithan Suvarna D, [2]Ms. Pavanalaxmi, [3]Mr. Savidhan Shetty C S

[1]UG Student, [2,3]Assistant Professor, Sahyadri College of Engineering & Management, Mangaluru, INDIA, [1]chaithan97@gmail.com, [2]pavanalaxmi.ec@sahyadri.edu.in, [3]savidhan.cs@gmail.com

Abstract: Lung cancer is one of the major threats faced by people around the world. Detection of lung cancer in the early stages can save lives. Human radiologists use CT scan images of the chest cavity to find the presence of tumor. This manual observation by the radiologists may involve human biases and errors which leads to false positives/false negatives. In order to overcome the human error involved in the diagnosis of lung cancer most of the research is done for the development of computer based diagnosis system. These systems make use of deep neural network which takes lung CT scan images as input. This project makes use of transfer learning a machine learning method which saves time and also uses limited computing power.

Keywords — Cancer Detection, Computational Power, Machine Learning, Neural Network, Radiologist, Transfer learning.

I. INTRODUCTION

The abnormal growth of cells causes cancer or tumor. The body part where these abnormal cells are present decides the type of cancer. This abnormal cell growth in lungs are called lung nodules, lung nodules are small masses of tissue in the lungs. The tumor can be malignant or benign. Benign tumors are noncancerous if they are of normal size they aren't harmful, they do not spread to other tissues or organs. If the benign tumor grows for some reason it can be dangerous. Malignant tumors are cancerous they spread to other parts of the body. The tumors appear as round, white shadow in the CT scan images. Lung nodules are usually 0.2 inch (5 millimeters) to 1.2 inches (30 millimeters) in size.

Based on the appearance of tumor cells in microscope, lung cancer can be classified as Small cell lung cancer (SCLC) and Non small cell lung cancer (NSCLC). These 2 types of lung cancer spreads and are treated in different way so making this distinction is important. SCLC account for about 10%-15% of lung cancer. This is the most aggressive type of lung cancer as it grows rapidly. SCLC is caused strongly due to cigarette smoking. SCLC rapidly grows to many parts in the body and is discovered after extensive spreading. SCLC is difficult to detect as the tumor cells will be very small and they resemble normal cells. NSCLC is the most common type of lung cancer accounting for 75-80% of the total lung cancer cases. Symptoms of Lung cancer:

1. Loss of appetite
2. Wheezing
3. Tiredness or weakness
4. Pain in the chest, shoulders, back
5. Cough producing blood
6. Swelling in the face and neck

Tests to diagnose lung cancer are as follows:

Imaging tests: The X-Ray image of a chest cavity can reveal abnormal mass of tissues in the lungs and CT scans can be used to determine small lesions in the lungs which cannot be detected by X-Ray.

Sputum Cytology: The patient who has cough and is producing sputum, the observation of the sputum under the microscope can reveal this cancerous cells in the lungs

Biopsy: The sample of abnormal cells in the lungs can be removed from the lungs and careful analysis in the lab can help in the detection of lung cancer

The human radiologists have to distinguish lung lesions from bones, pulmonary veins and other complex structures in the lung cavity this may lead to errors in diagnosis of lung cancer. The error in diagnosis of lung cancer can lead to legal issues and also unnecessary medical expenses. The observer may overlook small lesions in lungs which are cancerous. In order to overcome the human error and biases a computer based diagnostic system can be designed which uses high-end image processing algorithms and deep convolutional networks that uses chest CT scan images to predict the occurrence of cancer.

II. LITERATURE SURVEY

The paper titled "Deep Learning for Categorization of Lung Cancer CT Images" by Allison M Rossetto and Wenjin Zhou Department of computer Science University of Massachusetts Lowell [1] shows the accuracy of the initial stage automated diagnosis of medical scans. They developed a convolution Neural Network (CNN) and also preprocessing pipeline to increase the accuracy of screening processes. It improves by clarifying both smoothed and unsmoothed images. The resulting predictions from each of

 DOI : 10.35291/2454-9150.2019.0260

the networks are then combined using a voting system designed. Lung cancer imaging Database provided by Kaggle Data Science Bowl 2017 which actually is a dataset with 1500patiebts and over 1,50,000 CT images. Both raw image data and images which are smoothened with Gaussian Filter as inputs to CNN is being used. They implemented two CNN ensembles in Matlab. One is Matlab Neural Network Toolbox and another without the toolbox in Matlab. Basic pipeline for both methods are same. Uniqueness in the paper is that since 2 methods are used, a voting system is done rather than taking average.

The paper titled "Study of Lung Cancer Detection by Image Processing" by Weixing Wang, Shuguang Wu Department of Computer Science and Technology Chongqing University of Posts and Telecommunication [2] suggest that decomposition algorithm and subtraction algorithm that has newer imaging and diagnostic methods and from the test results, proposed techniques were successful to detect tiny spots on CT image. To apply image processing techniques into lung tissue information recognition, newly developed ridge detection Algorithm is applied here. Algorithm is been compared to traditional image segmentation Algorithm. All the result was found satisfactory. Before considering in detail the morphological information that is available from images of the early lung cancer, it is useful to implement the image techniques that have ever been used to quantify disease extent. The aim of detecting in this paper is to auto-tracing early lung cancer tissue, which is one of the most difficult tasks in detecting target.

The paper titled "Deep Learning Application Trial to Lung Cancer Diagnosis for MedicalSensor System" by Ryota Shimizu, ShusukeYanagawa, Yasutaka Monde, Hiroki Yamagishi, Mototsugu Hamada, Toru Shimizu, and Tadahiro Kuroda Faculty of Science and Technology Keio University Yokohama, Japan [3] tried to apply Deep Learning to human urine data and achieved 90 % accuracy in the determination of lung cancer.

They also proved that Deep Learning is also effective for human vital data analysis and we can do pre-diagnosis without any special medical knowledge.Health diagnosis is very difficult because it needs knowledge or experience of medical science but pre diagnosis via sensing data may be easy to do, for example temperature of body is a sensing data generated by thermometer, we know that high temperature is a sign of fever. Like our major goal is to build an easy health checking system which can detect some diseases by analysing sensing data such as breath, saliva, urine, which can be collected without hurting human body. Deep neural network or Deep learning is a well known method of machine learning and it is effective for feature extraction from pictures. Then they thought deep learning also can extract features from sensing data, here in this case they tried to build a diagnosis system of lung cancer based on deep learning. Input data of system was generated from human urine by Gas Chromatography Mass Spectrometer. Gas Chromatography Mass Spectrometer is an instrumental technique, comprising a gas chromatograph coupled to a mass spectrometer by which complex mixture of chemical may be separated, identified and quantified.

The paper titled "Segmentation of Sputum Cell Image for Early Lung Cancer Detection by N. Werghi, C Donner, F. Taher, H. Alahmad Khalifa University, UAE, University Osnabrueck Germany [4]. This method exhibits an elegant and methodological choice of threshold parameter. They found that sputum cell segmentation mean shift technique significantly outperforms the Hopfield Neural Network (HNN) technique. Framework for detection and segmentation of Sputum Cells in Sputum Cells in sputum images. It is compared with various experiments with data set of 88 images. Sputum cytology is a method for early lung cancer detection. In this the physicians examine a sample of sputum collected from a person's mucus under microscope. Goal of this research was to design a computer aided diagnostic system of the sputum stained smears. There were two challenges to overcome one is to detect sputum in smeared solution and one more is segmentation of sputum cells into cytoplasm and nucleus. For detection of sputum cell they used Bayesian classification framework. For sputum cell segmentation they used the application of robust mean shift technique.

The paper titled "Analysis of statistical texture features for automatic lung cancer detection in PET/CT images" by K. Punithavathy Research Scholar, Department of ECE, Hindustan University M. M. Ramya Professor, Centre for Automation & Robotics, Hindustan University. Sumathi Poobal Professor, Department of ECE, KCG College of Technology [5] successfully developed automatic lung cancer detection for PET/CT images using texture analysis and FCM. This paper aims at developing a methodology for automatic detection of lung cancer from PET/CT images. Image pre-processing methods such as Contrast Limited Adaptive Histogram Equalization (CLAHE) and Wiener filtering were performed to remove any errors due to contrast variations and noise. Lung region of interest (ROI) were extracted from images using morphological operators for which Haralick statistical texture features were preferred as they extract more texture information. Fuzzy C means (FCM) clustering was used to classify the regions as normal or abnormal. The proposed method was carried out using PET/CT images of lung cancer patients and implemented using MATLAB. The performance of the proposed methodology was evaluated using Receiver Operating Characteristics (ROC) curve. The proposed method provides better classification and accurate cancer detection.

 DOI : 10.35291/2454-9150.2019.0260

Image pre-processing methods were used to enhance the images.

Methodology for the proposed method as follows

1) Pre-Processing: Pre-processing techniques such as CLAHE and wiener filtering were applied to reduce the artifacts due to contrast variations and noise without affecting the image details.
2) ROI Extraction: For accurate lung cancer detection, it is necessary to accurately extract lung regions from the surrounding anatomical parts. Using Morphological Closing operations, except the Lung Lobes, all the external structures and the internal parts of lung like blood vessels, bronchi were eliminated from the binary images.
3) Feature Extraction: Medical images contain more texture information. The presence of lung cancer drastically changes the appearance of the texture of the lung In Medical Imaging, extracting the texture features for automatic differentiating between normal and abnormal tissues is of main importance.
4) Classification: Classification of PET/CT lung images is required to identify the presence of lung cancer. Here, FCM algorithm is preferred for classifying the regions as normal or abnormal (cancer).

III. METHODOLOGY

We are using Anaconda distribution which is a free open source distribution of python/R programming languages for Data science and machine learning applications. This simplifies package management as it comes with pre-contained packages/libraries like numpy, scipy, matplotlib, tkinteretc and also IDEs like spyder and browser based IDE like Jupyter notebook. We also use Keras which is a high level neural network Application Programming interface (API) which has several inbuilt classes and functions which can be used to implement complex deep neural network. We have also designed a GUI which makes the system user friendly. We have used python's tkinter library to create the GUI

Keras is a open source python library which can be used to implement high end neural networks. Keras runs on top of tensorflow/theano or CNTK faster implementation and experimentation with neural networks. Keras uses tensorflow by default in the backend; here backend refers to some low level computations. It uses higher level set of abstractions which makes it easy to develop deep learning networks regardless of the computation used in the backend. Keras also provides numerous implementations of basic building blocks of neural networks such as layers, activation functions, loss functions and optimizers. The codes are hosted on GitHub and it can be accessed by anyone without any monetary charges.

Keras also supports convolutional neural networks (CNN) and Recurrent Neural networks (RNN) by providing implementations for different models like Sequential and also provides layers like convolution, Pooling, Dense etc. Keras is user friendly, easily extensible, modular and works with python. The core data structure in keras is model which is used to organize several layers. The simplest type of model is Sequential which is a linear stack of model. We can add layers to this model by using .add() command.

Anaconda is a package and environment manager, it also acts as a python distribution which provides over 2000 plus packages and libraries. Anaconda distribution can be downloaded from its official website, which has several version of Anaconda distribution. After downloading it from the Anaconda official website we can install it on our system. Anaconda distribution is available for windows, Linux and Mac. Anaconda also comes with Anaconda navigator which is desktop GUI which can be used to access its features and Anaconda prompt which is a command prompt in window and terminal in Linux and Mac. We can install, update and remove any package from Anaconda by few clicks in Anaconda navigator or by writing some command in the Anaconda prompt. Anaconda comes with over 200+ packages preinstalled and over 2000 open source packages can be installed later as per our need. We can download and install packages from Anaconda repository using the command "conda install" in the Anaconda prompt and many other packages can be installed using the command "pip install". We can make our own custom packages and upload it to Anaconda cloud using the command "conda build". Anaconda also comes with Spyder IDE and also Jupyter Notebook.

Spyder is a scientific environment written in python for python to be used by scientists, engineers and data scientists. It provides the functionality of editing, data analysis, profiling and debugging for a comprehensive development tool with data exploration, deep inspection and interactive execution and visualization capabilities.

Jupyter notebook is a open source web application, it can be defined as a web based IDE. It allows us to create and share documents which contain live code, visualization, mathematical equations and plain text used to describe the code. It can be used for data cleaning and transformation, statistical modeling, numerical visualization and Machine learning. Jupyter notebook supports for over 40 programming languages which includes python, Scala, R and Julia etc. Notebooks and also be shared with others through Dropbox, Gmail and Github. Jupyter Notebook provides interactive output i.e. the codes are written within

 DOI : 10.35291/2454-9150.2019.0260

cells and each cell can be run individually which displays output for a given cell.

Training a Neural network from scratch requires large dataset and also training time will be high. Instead of doing this we can use transfer learning where we transfer the weights which is already learned by the network and use it to solve the target problem. Transfer learning is a machine learning methodology where knowledge gained when solving one problem is stored and it is applied to some other related problem. The knowledge here is nothing but a model which was developed for that problem along with the pretrained weights. Here we use models created by others as starting point to create model for our problem by making changes to the chosen model. Transfer learning is used for computer vision and natural language processing applications which requires heavy computing power. The main idea here is to use pretrained models which are previously trained on large dataset using high end GPUs.

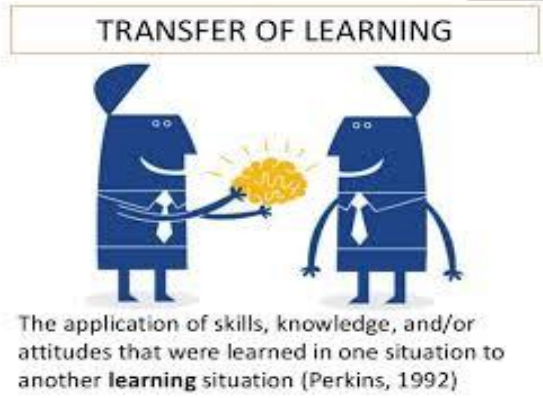

Fig 3.1 Transfer learning intuition

A. *Pretrained models*

Pretrained models are models which are developed by someone else to solve some problem using large dataset and high computing power. Keras library has this pretrained models built in which comes with the weights and biases. These pretrained models are previously trained network, which is trained on large dataset to be used in multi class classification. Some of the commonly available pretrained models are VGG16, VGG19, Inception V3, Mobilenet, Resnet 50 etc. Pretrained models may not be 100% accurate but it saves lot of time required for training.

This pretrained model can be used in an application by making some changes to it. Either architecture of the pretrained model can be choosen by initializing all the weights randomly or training the model from scratch using the dataset. This pretrained model can be used as a feature extractor by removing the output layer; it could extract some basic features like edges, curves, shapes and lines etc. User can also partially train the model by freezing some layers and fine tuning the others. Many pretrained model architecture are readily available on Keras library. These pretrained models are trained on Imagenet which is a database of millions of images. These pretrained architectures are capable of classifying 1000 different categories. The convolution layers in the pretrained models are used for feature extraction and initial layers in the pretrained models are composed of convolutional layers. These convolutional layers can be used to extract features in the above application, so this layer is freezed so that same weights can be used for extracting features. These layers are capable of detecting edges and shapes so the same layers can be utilized for feature extraction in cancer detection.

B. *VGG16 pretrained model*

VGG16 is a convolutional neural network architecture which was proposed by K. Simonyan and A. Zisserman in the paper “Very Deep Convolutional Networks for Large-Scale Image Recognition”. This model achieves the accuracy of 92.7% on ImageNet which is a dataset of over 14 million images belonging to thousand classes. It uses multiple filters of size 3X3 one after another. VGG16 was trained using Nvidia Titan Black GPU and it was trained for weeks on ImageNet dataset. ImageNet is a database of about 15 million high resolution images belonging to 22,000 different categories. These images are collected from online and labeled by human labellers using Amazon’s Mechanical Turk crowd-sourcing tool. VGG-16 is named so because it was developed by Visual Geometry Group (VGG) which is a research group in Oxford University. The number 16 refers to the number of trainable layers. It has total of 138,357,544 parameters and all parameters are trainable. VGG16 takes a input image of fixed size i.e. 224X224 RGB image. Here image is passed through stack of convolutional layers with a filter of size 3X3 and convolutional stride is fixed to one pixel. Maxpooling layers is used to perform spatial pooling which follows some of the convolutional layers it uses a window of size 2X2 with stride two.

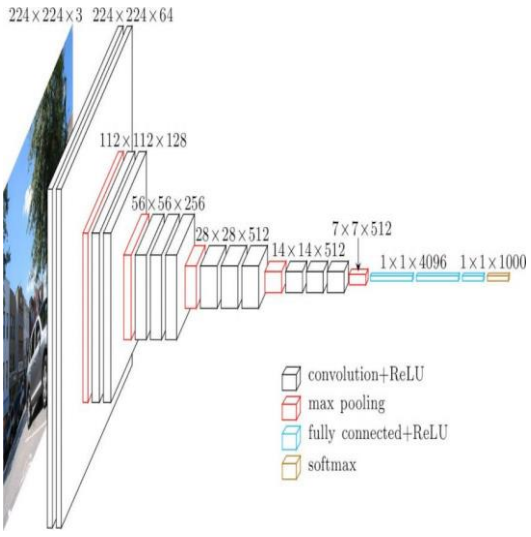

Fig 3.2 VGG16 architecture

 DOI : 10.35291/2454-9150.2019.0260

The stack of convolutional layers are followed by 3 fully connected layers, the final fully connected layer acts as a output layer which can classify 1000 classes hence it has 1000 channels. The Pooling and flattening layers does not learn anything so it has 0 trainable parameters. The pooling layer has a window size of 2X2 with stride 2; hence it outputs a pixel for every 4 pixels and jumps by 2 pixels to do the next calculations.

The VGG16 model is pretrained on large image dataset we can use the pretrained weights which are learned during the training process. The initial few layers of vgg16 model isused for detecting the basic features, middle conv layers are used for detection of some shapes and top layers are used for some complex feature extraction and classification. We will freeze all the layers in vgg16 except the last 3 layers so that the freezed layers won't be retrained and the pretrained weights can be retained. The output layer in vgg16 can be removed since it has output shape of 1000 i.e. it can classify 1000 classes. In our application of lung cancer detection we only want to classify images belonging to 2 classes we can replace the output layer with the dense layer of shape 2 which can be used for classifying the images into cancerous and non cancerous. Since we are freezing only the bottom layers and fine tuning the top layers we do not need large dataset and also training time will also be less compared to training the CNN from scratch.

The Lung cancer detector is a system which is used to detect whether the patient suffers from lung cancer or not by detecting the presence or absence of tumors in the lung CT scan images. To make the system user friendly we have designed a GUI which can be used to select the name of the image from the pool of images and use it to predict the output. The GUI is designed using a python library called tkinter, it is the fastest way to create GUI applications.
The GUI features 4 different windows

1. lung cancer detector (Home window)
2. upload window
3. prediction window
4. About window

C. Main window

Home window is the first window which opens when we run the code for GUI. It has 2 buttons which are labeled as 'upload' and 'About'. The upload button directs the user to the upload page and about button opens About window.

Fig 3.3 Home window

D. Upload window

Upload window features a Combo box using which we can select the names of images which can be used for prediction the window also contains a button labeled 'predict' which directs the user to the prediction window.

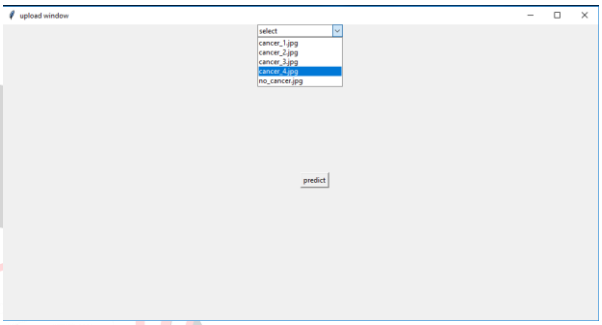

Fig. 3.4 Upload window

E. Prediction window

Prediction window displays the output whether the image selected is cancerous or non cancerous. It has 2 buttons labeled 'Home' and 'Back' which takes to Home window and previous window respectively.

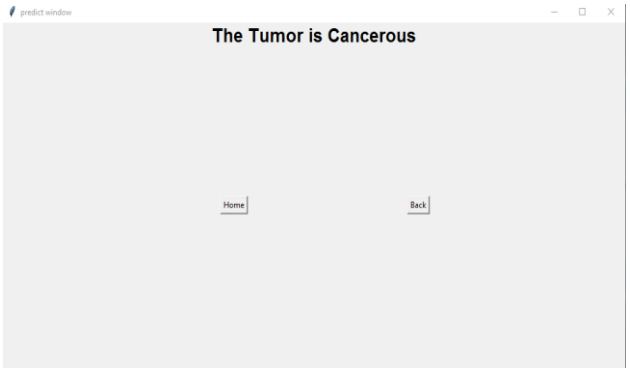

Fig. 3.5 predict window

IV. CONCLUSION

We have studied and analyzed the different methods of diagnosis of the lung cancer detection. Application of

 DOI : 10.35291/2454-9150.2019.0260

computer based lung cancer detection system to help doctors to make better and informed decision when diagnosing the lung cancer. We have studied different research papers in the field of lung cancer detection using machine learning, image processing, deep learning; we have explored various software tools in python, libraries which are used to implement machine learning algorithms. We have implemented a Convolutional Neural network using a VGG16 pretrained model which can be trained to classify the lung CT scan images into cancerous and non cancerous.

REFERENCES

[1] Allison M Rossetto and Wenjin Zhou, "Deep Learning for Categorization of Lung Cancer CT Images", Department of computer science University of Massachusetts Lowell.

[2] Weixing Wang, Shuguang Wu, "Study of Lung Cancer Detection by Image Processing", Department Of Computer Science and Technology Chongqing University of Posts and Telecommunication.

[3] Ryota Shimizu, ShusukeYanagawa, Yasutaka Monde, Hiroki Yamagishi, Mototsugu Hamada, Toru Shimizu, and Tadahiro Kuroda, "Deep Learning Application Trial to Lung Cancer Diagnosis for MedicalSensor System", Kuroda Faculty of Science and Technology Keio University Yokohama, Japan.

[4] N.Werghi, C Donner, F.Taher, H.AlahmadKhalifa, "Segmentation of Sputum Cell Image for Early Lung Cancer Detection", University Osnabrueck Germany.

[5] K.Punithavathy,M.M.Ramya, "Analysis of statistical texture features for automatic lung cancer detection in PET/CT images"

[6] R. Fakoor, F. Ladhak, A. Nazi, and M. Huber, "Using deep learning to enhance cancer diagnosis and classification," in Proceedings of the International Conference on Machine Learning, 2013.

[7] H. Yang, H. Yu, and G. Wang, "Deep learning for the classification of lung nodules," arXiv preprint arXiv: 1611.06651, 2016.

[8] Matsumoto M, Horikoshi H, Moteki T, et al. A pilot study with lungcancer screening CT (LSCT) at the secondary screening for lung cancer detection. Nippon Acta Radiol 1995; 55:172- 17.

[9] Kaneko M, Eguchi K, Ohmatsu H, et al. "Peripheral lung cancer: screening and detection with low-dose spiral CT versus radiography." Radiology 1996; 201:798-802.

[10] Muxuan Liang, Zhizhong Li, Ting Chen, and Jianyang ZengIntegrative "Data Analysis of Multi-platform Cancer Data with a Multimodal Deep Learning Approach," IEEE/ACM Transactions on Computational Biology and Bioinformatics, pp. 928-937, Feb. 2011.

[11] Dignam JJ, Huang L, Ries L, Reichman M, Mariotto A, Feuer E. "Estimating cancer statistic and other-cause mortality in clinical trial and population-based cancer registry cohorts", Cancer 10, Aug 2009.

 DOI : 10.35291/2454-9150.2019.0260