RESEARCH ARTICLE OPEN ACCESS

Effective Lung Cancer Cell Detection using Deep Convolutional Neural Networks

Swapnil Nivendkar*, Shreyas Satardekar**, Shreyank Prabhu***
**(Software Engineer, Samsung Research and Development, Bengaluru,*
***(Technical Analyst, Credit Suisse, Pune,*
**** (MS in MIS at Texas A&M University, Texas,*
Corresponding Author : Swapnil Nivendkar

ABSTRACT

With oncologists relying increasingly on low-dose CT scans to detect lung cancer, our study proposes a machine learning approach for early detection of Lung Cancer. While existing algorithms in the medical imaging domain focus on segmentation and diagnosis through traditional image processing techniques, we approach the problem by using Deep Learning technique called Convolutional Neural Networks. Using Google Cloud Engine for Machine Learning for training data overcomes the problems encountered while handling large scale data. The traditional methods take a lot of time before the doctor can finally tell the patient the result which in certain cases proves to be fatal. Our approach aims to provide the result of the tests at a much faster speed, thereby bringing up the survival rate of those that have lung cancer and decrease the mortality rate of Lung Cancer related deaths.

Keywords – Cloud, Lung Cancer, Neural Networks.

Date Of Submission:25-08-2018 Date Of Acceptance: 08-09-2018

I. INTRODUCTION

Lung Cancer is one of the leading causes of cancer related deaths [1]. Thus, it is important to be able to detect cancer in the lungs as early as possible. Small masses of tissues found in the lungs, known as pulmonary nodules, possess the risk of becoming cancerous. Therefore, being able to identify nodules is absolutely necessary to diagnose lung cancer in its early stages. These nodules, however, are difficult to detect, as they can be as small as 1-2mm.

The overall 5-year survival rate of lung cancer is poor (only 5%) [2]. The age of presentation in Indians is also younger (mid 50’s) while in the rest of world it presents in people in their mid-60’s. With such a high death rate it is better to avoid it. Quick diagnosis and prognosis of a type have become necessary in cancer research, which helps in further treatment of patients.

In our system, we focus on Convolutional Neural Network (CNN) for the development of predictive models, resulting in effective and accurate decision making. Machine learning will always improve the way we detect cancer progression however accurate ways of validation are needed in everyday clinical use.

II. INDIAN SCENARIO

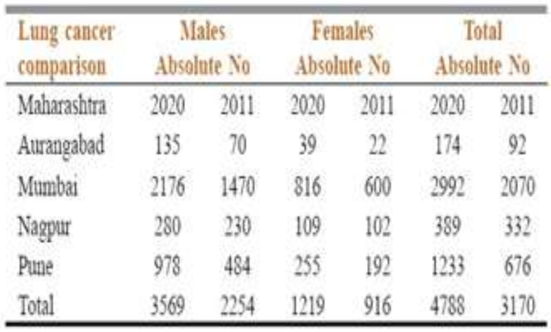
Smoking tobacco, cigarettes being one of the most prominent factors causing lung cancer is evident in Indian men however in the case of Indian women, the association of smoking is not so strong, which points out to other factors that possess the risk of causing lung cancer [5]. Despite the progressive changes in the diagnostic methods, molecular changes and therapeutic interventions, the outcomes of lung cancer remain poor, therefore a better knowledge of the risk factors will assist us to develop preventive measures at the community level.

The statistics related to lung cancer paint a picture of the scenario in India, while 6.9 per cent of all new cancer cases are lung cancer, it constitutes 9.3 per cent of all cancer related deaths in India. Highest reported deaths are in Mizoram in both sexes. Lung cancer is the most common cancer type in men [4]. The patter of lung cancer differs by various factors such as ethnicity, geographic region and highlights the prevalence of smoking. 5-year survival rate is quite low being 15 percent in developed countries and 5 percent in developing. CT screening in high risk population showed a relative risk reduction of 20 per cent in lung cancer mortality but false positives of 96 per cent [3]. Dependability on such a screening tool in India is doubtful. Need for new non-invasive methods for quick and early diagnosis and screening of high risk population is of high priority in healthcare.

We lack the understanding presently of the changing epidemiological trends of lung cancer among Indian patients. Addition to that is the lack of understanding in the alarming increase of lung cancer amongst the nonsmokers. India is consistent with the global trend of adenocarcinoma.

On a microlevel we have limited understanding of the impact of the factors that are specific and vary from region to region such as the presence of indoor air pollutants, the use of domestic or biomass fuel exposure, the presence of lack of micronutrients in our diet, and the possible contribution of infectious pathogens such as Mycobacterium tuberculosis.

The projected change in incidence of lung cancer in Maharashtra by 2020 which is only 2 years away can be seen in the Fig.1 below. In 4 cities of Maharashtra (Mumbai, Pune, Nagpur and Aurangabad) the absolute increase from 2016 to 2020 can be seen to be increasing from 3170 to 4788(more than 50% increase) [6]. At a national level this would create about 235,104 new patients, with about 90% of these in an advanced inoperable stage, the future looks extremely challenging.



Lung cancer comparison	Males Absolute No		Females Absolute No		Total Absolute No	
Maharashtra	2020	2011	2020	2011	2020	2011
Aurangabad	135	70	39	22	174	92
Mumbai	2176	1470	816	600	2992	2070
Nagpur	280	230	109	102	389	332
Pune	978	484	255	192	1233	676
Total	3569	2254	1219	916	4788	3170

Fig. 1 – Projected change in incidence of Lung Cancer.

III. GENERIC LUNG CANCER DETECTION METHODS.

3.1 Imaging Tests

An X-ray image of your lungs may reveal an affected mass or nodule which might not appear normal on imaging. A CT scan may reveal minor differences from an X-ray by displaying small lesions which might be missing from an X-ray.

3.1.1 CT Scan

A CT Scan can be used by doctors to see the location as well as the size of the tumor. A 3D image is generated of the inside of the body using X-Rays. The Computer then combines these images into a cross-sectional view to point out any abnormalities.

3.1.2 Positron emission tomography (PET) scan

This is usually combined with a CT scan. This is used to create images of organs inside our body. The most important drawback of this technique is that a small amount of radioactive substance is injected into the patient's body. A scanner then detects this substance to produce images to display insides of one's body.

3.1.3 Magnetic resonance imaging (MRI) scan

Detection of location of lung cancer can be done using an MRI scan. Magnetic Rays and not x-rays are used to generate the detailed images of the body. A Special dye called a contrast medium is injected into the patient's vein or is given as a pill to follow. The reason why MRI scans are not widely used for Lung Cancer Detection is that it does not work well to take pictures of parts of the body that are moving, like your lungs, which move with each breath you take. Thus, it is not generally used for detection of Lung Cancer.

3.2 Sputum Cytology

Sputum cytology examines a sample of sputum to find the presence of abnormal cells. Sputum is produced in the lungs and the airways and is not as same as saliva, it may have normal lung cells in it. Sputum is collected by coughing up mucus or by breathing in a saltwater mist and then coughing [7].

It may take several days to receive the results of sputum cytology; therefore, it is a time-consuming process. Depending on the sputum constituents the accuracy may vary of the cytology and therefore possibility of false negatives may increase.

3.3 Needle Biopsy

A lung needle biopsy is used to diagnose an irregular area of tissue in the lungs. This procedure is used to obtain a small sample of the lung tissue. It is also referred to as percutaneous needle aspiration. This procedure is recommended by the doctor to check whether the lung mass is benign or malignant, the stage of the malignant tumor, to explain why fluid has collected in the lung cavity, etc. However, this method has a number of disadvantages such as the patient may suffer from excessive bleeding, infection in the lungs, the patient might also suffer serious complications such as collapsing of the lungs and coughing up blood [8]. These problems are further worsened since they might persist for a while after the procedure has been performed. The doctor instructs the patient when either of the following symptoms show up:

1) Bleeding from the autopsy site
2) Coughing up more blood than a small amount of blood
3) Severe chest pain

4) Redness or drainage at the biopsy site
Along with these disadvantages, the patient cannot perform daily routine activities for quite a while until your doctor instruct you to resume your routine.

IV. CONVNETS IN MEDICAL IMAGING

4.1 Generic Convolutional Neural Network

Convolutional Neural Networks (**ConvNets** or **CNNs**) are a category of Neural Networks that are effective in image recognition and classifying images [11]. CNNs consist of multiple layers of reception fields. These are small neuron like collections which process portions of the input image. Output collections are tiled with input so that they overlap, which obtains a higher-resolution representation of the original image; this is repeated for every such layer. Various combinations of convolutional and fully connected layers are present in this model. On small regions of input convolutional operations are brought in to reduce the number of free parameters and improve generalization [12].

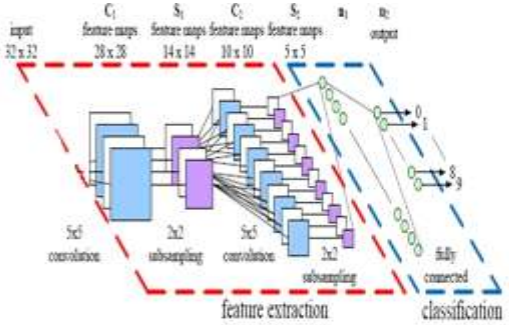


Fig 2 – Layers in Convolutional Neural Network.

4.2 Dataset

The Lung Image Database Consortium and Infectious Disease Research Institute (LIDC/IDRI) contains an image collection of diagnostic and lung cancer screening thoracic CT scans [9]. It consists of 888 CT scans with marked-up lesions that we use as ground-truth labels for the classification problem. Consequently, the annotations file additionally contains false positives that we can incorporate into our training set. The computerized developed nodules were marked by different radiologists resulting in a very unbalanced dataset of mostly false positives. [10].

4.3 Image Extraction

Each of the 888 CT scans consists of a MetaHeader (.mhd) file and the unprocessed multidimensional scan in a raw format. A significant amount of time had to be invested in generating the 2-dimensional images of potential nodules by translating the coordinates in the mark-up file to select the correct cross-sectional slices of the lung scan, crop them, and store them in a traditional image file format. Using the Simple Insight Segmentation and Registration Toolkit (SimpleITK), we read the raw scan in and converted it into a 3-dimensional array. The annotations file has the candidate locations in world coordinates which need to be converted to non-integer voxel coordinates to correctly identify the region in the array containing the potential lesion. Now 50 x 50 grayscale images were generated for training, testing and validating a CNN. As a result of above process, we had under-sampled the negative class such that every 1 in 6 images had a nodule. Further Augmentation was carried out on these images which resulted in an ideal 85-15 class distribution. The organization running the LUNA challenge responsible for the dataset split the dataset into roughly 80% training, 10% validation and 10% test set [10].

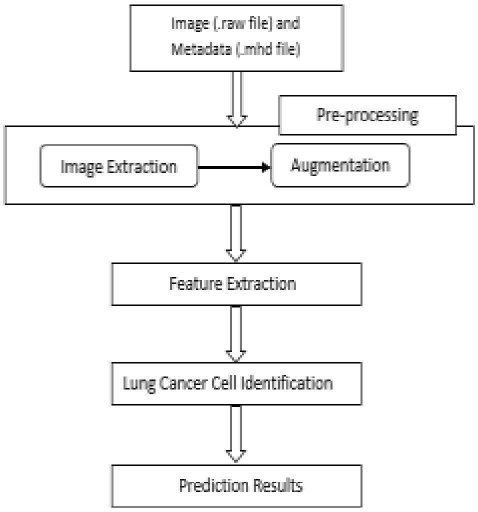


Figure 3 – Diagnosis Process

4.4 Applying CNN

As we pass an image through the first convolutional layer (50 x 50 x 32), it generates a feature map.
The max pooling layer following the first layer down-sampled the feature map by 2. Similarly feature maps were generated by third convolutional layer.
We had a validation accuracy of 93 %. Our model has a precision of 89.3 % and recall of 71.2 %. The model has a specificity of 98.2 %.

V. INCORPORATING CLOUD WITH IMAGING

We plan to use the cloud services since we will be handling a very large dataset. Having data storage and processing hardware facility will incur a high initial cost. Also, additional effort would have to be taken to manage such facilities proving to be expensive in the long run as well. There are a number of cloud-based services out in the market

such as Google cloud service, Microsoft Azure, and the Amazon Web Services, etc.
The reason we chose Google Cloud is for the numerous features it provides which are discussed below. Also, the cost of using the Google cloud services is nominal though it increases according to the volume of data we will dealing with. This puts light on one of the features of the service that is scalability. If the traffic increases i.e. to say that the cloud service gets more data, then its service can be expanded for additional cost to handle the traffic. In the following section we will learn more about the Google cloud service and how it can be implemented for machine learning of the Lung Cancer dataset.

5.1 About the Google Cloud Platform

It is a cloud computing service widely used by google. It provides developer products to build large range of programs. Its advantage is that it provides a set of modular cloud-based service with host development tools, such as hosting, computing & data storage.

5.2 Features of Google Cloud Service

5.2.1 Managed Scalable Machine Learning

Google Cloud Machine Learning Engine allows you to develop models which work on any kind of data.

5.2.2 Predictive Analysis at scale

Integration with Google global load balancing allows automatic scaling to reach users world wide

5.2.3 Fully managed service

GPU based acceleration, managed by google allowing you to focus and models and not clusters.

5.2.4 Efficient deep learning capabilities through TensorFlow model

The main idea here is to use TensorFlow to train our data. However, for a very large dataset where data consists of CT scans from all over India, it is very difficult, time consuming and virtually impossible to train such large data. Therefore, we will make use of the Google Cloud Service to store our data in the cloud and to train it using TensorFlow on the powerful machines on the google cloud. TensorFlow is an open source software library for ML and developed by Google for building and training NNs. While the reference implementation runs on single devices, TensorFlow can run on multiple CPUs and GPUs (with optional CUDA extensions for general-purpose computing on graphics processing units).The Google cloud shell is based on Debian with all the necessary tools installed.

5.3 Initializing Machine Learning on Cloud

We start the Google Cloud shell to run the following scripts.

1) Additional tools are installed with the help of the following commands

```
“curl
https://storage.googlepis.com/cloudml/scripts/setupcloudshll.sh|bash”
```

2) Adding the tools to the path with the following script

```
“export PATH=${HOME}/.local/bin:${path}”
```

3) The environment is then verified using the following script

```
“curl
https://storage.googlepis.com/cloudml/scripts/setup_cloud_shell.sh |python”
```

4) Next the cloud ML project is initialised

```
gcloud beta ml init-project
```

To have access to cloud storage we need to setup the Google Cloud Storage Bucket. The cloud Ml services need to access cloud storage locations to read and write data during model training and batch prediction. The following section illustrates how to create Google cloud bucket for reading and writing data during model training and batch prediction:

1) We first set a name for the bucket.

```
PROJECT_ID=$(gcloud config list project –format “value(core.project)”)
BUCKET_NAME=$(PROJECT_ID)-ml
```

2) Next the new bucket is created.

```
Gsutil mb us-central1 gs://$BUCKET_NAME
```

With the above steps we have installed two Google cloud commands Gcloud which is a command line interface to Google Cloud Platform, it includes the following:

1) Google cloud machine learning jobs
2)Google compute engine virtual machine instances and other resources
3) Google cloud dataproc clusters and jobs
4) Google cloud deployment manager deployments
5) Gsutil, a command line interface for Google cloud

To execute our data in the cloud we will have to submit our code to the Google cloud which is done with the help of a STAGING_BUCKET.

1) Select a job-name and then in the next command we choose the staging bucket where our source code will be stored.

```
JOB_NAME=”LungCancerDetection"
PROJECT_ID=’gcloud config list project –format
“value(core.project)”’
    STAGING_BUCKET=gs://$(PROJECT_ID)-ml
```

2) Then the following commands are executed in the Google command line

```
 $gcloud beta ml jobs submit training
        ${LungCancerDetection) \
> --package-path-train \
> --staging-bucket="${STAGING_BUCKET}"\
> --module-name-pythonFileName
```

3) To copy the data that needs to be trained from our local machine to the cloud we make use of the following command

```
INPUT_PATH=${STAGING_BUCKET}/input
gsutil cp input/input.csv $INPUT_PATH/input.csv
```

Another easier way which involves the GUI of Google cloud is to open the Google cloud menu and select the storage option, then we select browser here we see all our uploaded files as well as options at the top to upload a new file, upload folder or create a folder, etc.

Additionally, the user can view all the jobs that he has done on the Google cloud machine learning platform as well as view the logs of each job which show the output of the job

With the saver function from TensorFlow we can create output directories to store the output of our computations. Similar to how we took input from files here for output we will use OUTPUT_PATH. The script for creating output directories is same as for inputting except for changing INPUT_PATH to OUTPUT_PATH. The output directory will contain the result and is stored in the cloud. The user can retrieve the file from the cloud as well.

The following figure gives an overview of how Google Cloud services are utilized in the Lung cancer machine learning project.

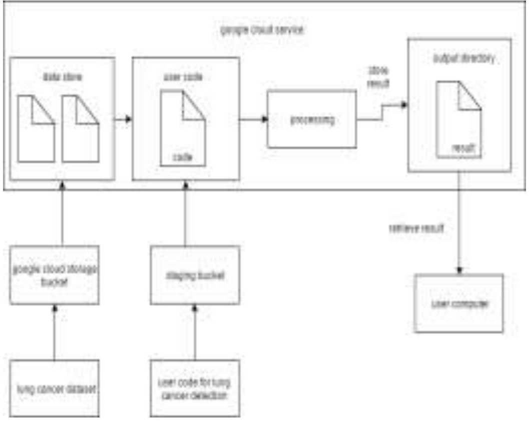


Fig. 4 – Cloud Services using Machine Learning

VI. CONCLUSION

We can consequently conclude that lung cancer is on the rise and cannot be determined based on smoking or non-smoking habits, a lot many factors contribute for cancer to prevail in one's body. A country where pollution is present in all forms, it keeps one susceptible to lung cancer. In such a scenario a quick system to detect lung cancer is a necessity, basing the possibility of cancer on the CT scan of a patient, the patient won't be wasting time finding a good oncologist and finding an appointment. Such systems deployed at hospitals might quicken the process and help the patient get aid immediately not allowing the cancer to progress in other stages thereby increasing the survival rate. Not only will this quicken the process but will also reduce the human error factor for example an incorrectly diagnosed CT scan, a small possibility where the doctor may miss the tumor, a highly trained machine may not.

CNN allows the CT scans to be fed as a data and train itself on basis of that. Trained by a large number of data increases the accuracy of such a network and increases the accuracy. A CT scan uploaded by the patient not only serves as a method for showing the possibility of lung cancer but also trains the network, with increasing no of CT scans fed to the network there will be an increase in accuracy.

REFERENCES

[1]. R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2015," CA: a cancer journal for clinicians, vol. 65, no. 1, pp. 5-29, 2015.

[2]. Dr. B. Jankharia Available: FTP: “http://picture-this.in/index.php/2016/02/04/lung-cancer-in-india/"

[3]. P. M. Parikh, A. A. Ranade. "Lung cancer in India: current status and promising strategies" in South Asian J Cancer. 2016 Jul-Sep; 5(3): 93-95. doi: 10.4103/2278-330X.187563.

[4]. Prabhat Singh Malik and Vinod Raina "Lung Cancer: prevalent trends & emerging concepts", 2015 Jan; 141(1): 5-7. pmcid: PMC4405940.

[5]. Vanita Noronha, Rakesh Pinninti, Vijay M. Patil, Amit Joshi, and Kumar Prabhash 'lung cancer Indian sub-continent' in South Asian J Cancer. 2016 Jul-Sep; 5(3): 95-103. doi: 10.4103/2278-330X.187571.

[6]. P.M. Parikh Available: FTP: “https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4991145/table/T2"

[7]. Webmd.com. [Online] Available FTP: “http://www.webmd.com/lung/sputum-cytology#1"

[8]. Webmd.com. [Online] Available FTP: “http://www.webmd.com/a-to-z-guides/fine-needle-aspiration#1"

[9]. The Lung Image Database Consortium and Infectious Disease Research Institute(LIDC/IDRI) [Online] Available FTP:

"https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI"

[10]. Consortium for Open Medical Image Computing [Online] Available FTP: "https://luna16.grand-challenge.org/data/"

[11]. Stanford.edu [Online] Available FTP: "http://cs231n.github.io/convolutional-networks/"

[12]. Ujjwal Karn [Online] Available FTP: "https://ujjwalkarn.me/2016/08/11/intuitive-explanation-convnets/"

Swapnil Nivendkar "Effective Lung Cancer Cell Detection using Deep Convolutional Neural Networks "International Journal of Engineering Research and Applications (IJERA) , vol. 8, no.9, 2018, pp 10-15