International Journal For Technological Research In Engineering
Volume 7, Issue 12, August-2020 *ISSN (Online): 2347 - 4718*

LUNG CANCER DETECTION USING IMAGE PROCESSING

Puneet Joshi[1], Priyadarshini M[2], Amogh K[3], Mr. Parashiva Murthy B M[4]
[1,2,3]B.E. Student, [4]Assistant Professor,
Dept. of C S & E, JSS S & T U Mysuru

***Abstract:** This study aims to highlight the significance of data analytics and machine learning in prognosis in health sciences, particularly in detecting life threatening and terminal diseases like cancer. Here, we consider lung cancer for our study. For this purpose, pre-existing lung cancer patients' data are collected to get the desired results. A predictive algorithm is developed to predict the probability of a patient catching lung cancer based on a dataset coming from the Data Science Bowl 2017. Data set (in the form of diagnostic images) is run past Matlab for analysis and forecasting. Image processing is employed for this purpose. Medical image segmentation and classification are done to achieve this. Classification depends on features extracted from the images. The emphasis is on the feature extraction stage to yield better classification performance. This information is then fed to machine learning algorithms to discern a pattern that can give some good insights into what combination of features are most likely to result in an abnormality.*

***Keywords:** MATLAB, Classification, Feature extraction, Image processing, Predictive algorithm*

I. INTRODUCTION

Lung cancer is one of the most common types of cancer, with nearly 225,000 new cases of the disease expected in the U.S. in 2016. Early detection is critical, as it opens a range of treatment options not available when cancer is detected at later, more advanced stages. Low- dose computed tomography (CT) is a potential breakthrough technology for early detection, with the ability to reduce deaths by 20%. Often, suspicious lesions identified in screening are initially assessed as high risk of cancer, but after additional follow-up tests, they turn out to be non-cancerous (false positives from the initial screening). Hopefully, machine learning can reduce the number of radiology exams flagged for potentially unnecessary follow up and avoid patient anxiety. Using a data set of high-resolution scans of lungs provided by the National Cancer Institute, researchers can develop artificial intelligence algorithms to accurately determine when lesions in the lungs are cancerous. This will dramatically reduce the false positive rate that prevents low-dose CT scans from being widely used for lung cancer detection.

II. PROPOSED SYSTEM

The proposed solution method for detecting lung cancer using CT scans of the patients is as follows:

- This project initially pre-processes the dataset in order to filter useless data and gets a smaller data set.
- In the pre-processing step, we can take into consideration the ratio of dark pixels to total number of pixels and select only 10 CT slices per patient which can yield us the proper features.
- Next, different image processing techniques are applied such as:
  - Bit-plane Slicing
  - Erosion
  - Median Filter
  - Dilation
  - Outlining
  - Lung Border Extraction
  - Flood-Fill Algorithm
- Using the above technique's, we make the image ready for feature extraction by removing the circle surrounding the lung region and also by highlighting the lung and the region inside the lung to get the nodule candidates.
- In the next step we are extracting three features namely, maximum drawable circle, area of candidate region, and mean intensity percentage.
- Then we are applying random forest classification to predict weather the extracted features contain a cancer nodule or not.

III. IMPLEMENTATION

a. Data Pre-processing

The raw data downloaded from Kaggle is quite large (about 150 GB) and it contains a lot of useless data. For data pre-processing, after reading the dcm format lung scan image into MATLAB we count the dark intensity in the specific lung region and compute a percentage threshold. With the threshold, we can filter a number of useless CT slices and finally pick 10 slices for each patient. As Figure shown, we do not want lung scans like CT3 but prefer lung scans link CT1, CT2 and CT3. After shrinking the data set, we obtain a smaller data set (about 4 GB) that can be processed in the next stage.

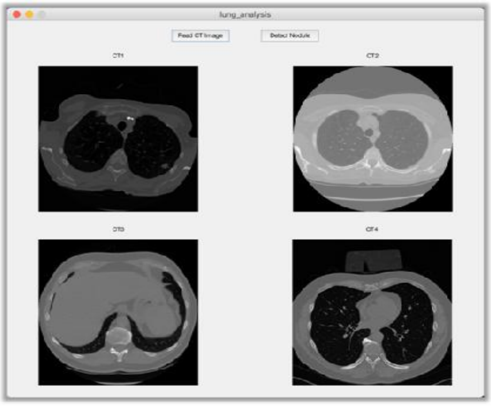

Figure 2. Some scan slices are useless

b. Detection Algorithm

i Lung Region Extraction

The initial stage of the proposed Computer Aided Diagnosing (CAD) [2, 7] techniques is the extraction of lung region from the CT scan image. The basic image processing techniques are utilized for this purpose. The image processing techniques applied in the proposed technique are Bit-Plane Slicing, Erosion, Median Filter,

Dilation, Outlining, Lung Border Extraction and Flood-Fill algorithms. Usually, the CT chest image not only contains the lung region, it also contains background, heart, liver and other organs areas. The main aim of this lung region extraction process is to detect the lung region and regions of interest (ROIs) from the CT scan image.

c. Segmentation of Lung Region

After the lung region is detected, the next process is segmentation of lung region in order to find the cancer nodules. This step will identify the region of interest (ROIs) which helps in determining the cancer region. In this project, Mean Shift is implemented for segmentation.

d. Analysis of Lung Region

After the segmentation is performed to the lung region, the features can be obtained from it and the diagnosis can be designed to exactly detect the cancer nodules in the lungs. The diagnosis rules can eliminate the false detection of cancer nodules resulted in segmentation and provides better diagnosis. The features that are used in this project in order to generate diagnosis rules are:

- Area of the candidate region
  With the help of this feature, the detected regions that do not have the chance to form cancer nodules are detected and can be eliminated. This helps in reducing the processing in the further steps and also reducing times taken by further steps.
- Maximum Drawable Circle (MDC)
  This feature is used to indicate the candidate regions with its maximum drawable circle (MDC). All the pixels inside the candidate region is considered as center point for drawing the circle. The obtained circle within the region is taken for consideration. Initially radius of the circle is chosen as one pixel and then the radius is incremented by one pixel every time until no circle can be drawn with that radius. Maximum drawable circle helps in the diagnostic procedure to remove more and more false positive cancerous candidates.
- Mean intensity percentage of the candidate region
  In this feature, the mean intensity percentage for the candidate region is calculated which helps in rejecting the further regions which does not indicate cancer nodule. The mean intensity percentage indicates the average intensity percentage of all the pixels that belong to the same region and is calculated using the formula:

$$\text{MeanPtg}(j) = \frac{\sum_{i=1}^{n} Intensity(i)/n}{Max}$$

where j characterizes the region index and range from the number of candidate regions in the whole image, which we specify as N. Intensity(i) indicates the CT intensity value of pixel i, and i ranges from 1 to n, where n is the total number of pixels belonging to region j. Max indicates the maximum intensity values of all of the candidate regions.

e. Training of Model

In this training section, we use the Random Forest model to train and test the data. The random forest starts with a standard machine learning technique called ‘Decision Tree’ which, in ensemble terms, corresponds to the weak learners. In a decision tree, an input is entered at the top and as it traverses down the tree the data gets bucketed into smaller and smaller sets. The random forest takes this notion to the next level by combining trees with the notion of ensemble. Thus, the ensemble terms, the trees are weak learners and random forest is a strong learner.

f. Model Testing and Evaluation

In this this project, we evaluate the random forest model with two methods:

- Cross-validation
  We test our classifier using a technique called “cross-validation”: train the classifier on all projects except for one. Here the projects mean the partition of the data set. Of course, we also know the ground truth for this held-out project, so we can see how well the classifier does on it, without cheating by training the classifier on the hold-out. We do this in turn with each project.

- Mean Precision
  The effectiveness of the classifier is the distance between the two means, which does not vary as threshold changes. One way of measuring the effectiveness of the classifier is the “precision”. Precision is the number of truly correct items (“hits”) divided by the number of items that the classifier says are correct (hits + false alarms). “Mean precision” takes into account the issues with choosing a threshold, noted above, by performing this calculation at a range of thresholds and taking the mean.

IV. EXPERIMENT RESULTS

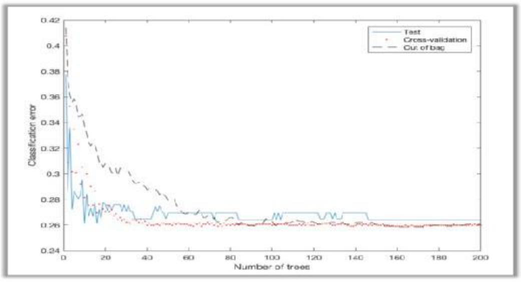

Figure 13. Cross validation and normal random forest model

As shown in figure, the following result presents a 10-fold cross-validated bagged ensemble and we examine the cross-validation loss as a function of the number of trees in the ensemble. With the number of trees increased by 50, the cross-validation model's classification error rate remains steady at 26%. While the normal random forest model's error rate fluctuates in the area between 26% and 28% when the number of trees increased from forty. The diagram also shows us that the mean accuracy of the normal random forest model is about 73%. With the help of trained models, we can compute a posterior probability of lung cancer for a patient automatically based on the detection algorithm depicted before.

V. CONCLUSION AND FUTURE WORK

This project presents the better Computer Aided Diagnosing (CAD) system for automatic detection of lung cancer. The initial process is lung region detection by applying basic image processing techniques such as Bit-Plane Slicing, Erosion, Median Filter, Dilation, Outlining, Lung Border Extraction and Flood-Fill algorithms to the CT scan images. After the lung region is detected, the segmentation is carried out with the help of Mean Shift clustering algorithm. With these, the features are extracted and the diagnosis rules are generated. These rules are then used for learning with the help of Random Forest. The experimentation is performed with 15, 000 images obtained from the kaggle contest. The experimental result shows that the proposed CAD system can able to tell the posterior probability of lung cancer for a patient based on the detection algorithm. Also the usage of Random Forest will increase the accuracy of detecting the cancer nodules. The futuristic scopes of this project are the following features:

- Using CNN to get better accuracy and precision.
- Creating an interface like a website or a mobile app.
- Integrating such applications with CT scan machines, so that along with the CT scans of the patient we will also get the result by the algorithm.
- Deploying this system in a real life scenario like a health care facility.

REFERENCES

[1] M.Gomathi,Dr.P.Thangarj, "A Computer Aided Diagonsis System For Detection of Lung Caner Nodules Using Extreme Learning Machine", ISSN:0975-5462, Vol. 2(10), 2010

[2] R. Wiemker, P.Rogalla, T. Blaffert, D. Sifri, O. Hay, Y. Srinivas and R. Truyen "Computer- aided detection (CAD) and volumetry of pulmonary nodules on high-resolution CT data", 2003. [3] D. Lin and C. Yan, "Lung nodules identification rules extraction with neural fuzzy network", IEEE, Neural Information Processing, vol. 4, 2002.

[3] S. G. Armato, M. L. Giger and H. MacMahon, "Automated detection of lung nodules in CT scans: Preliminary results", Med. Phys., Vol. 28, pp. 1552–1561, 2001.

[4] B.V. Ginneken, B. M. Romeny and M. A. Viergever, "Computer-aided diagnosis in chest radiography: a survey", IEEE, transactions on medical imaging, vol. 20, no. 12, 2001.

[5] M. Fiebich, D. Wormanns and W. Heindel, "Improvement of method for computer- assisted detection of pulmonary nodules in CT of the chest", Proc SPIE Medical Imaging Conference, vol. 4322, pp. 702–709, 2001.

[6] M. N. Gurcan, B. Sahiner, N. Petrick, H. Chan, E. A. Kazerooni, P. N. Cascade and L. Hadjiiski, "Lung nodule detection on thoracic computed tomography images: preliminary evaluation of a computer-aided diagnosis system", Medical Physics, vol. 29, no. 11, pp. 2552- 2558, 2002.

[7] R. Wiemker, P. Rogalla and R. Zwartkruis, T. Blaffert, "Computer aided lung nodule detection on high resolution CT data", Medical Imaging, Image Processing, Proceedings of SPIE, vol. 4684, 2002.