Lung Cancer Detection in CT Scans of Patients Using Image Processing and Machine Learning Technique

Karan Sharma[1](✉), Harshil Soni[1], and Kushika Agarwal[2]

[1] System Level Solutions (I) Pvt. Ltd., V.U. Nagar 388121, Gujarat, India
{karan.sharma,hsoni}@slscorp.com
[2] Department of CSE, Sikkim Manipal Institute of Technology, Sikkim, India
kushikaagarwal@gmail.com

1 Introduction

Lung cancer strikes more than 1.8 million people every year and accounts for billions of dollars in healthcare costs [1]. Early detection is critical to give patients the best chance of recovery and survival. All accepted contemporary works and studies have suggested methods to improve detection of cancerous lesions if any present in lungs. Many of the existing concepts have described about feature extraction methods for training of machines [2], but a huge void exists in terms of processing and analysis of such voluminous data by using simple neural network and other machine learning algorithms like random forest, decision tree, etc. Also, to note that images occupy comparatively large space and again to use any data stored in very low level format like JPEG will result in loss of crucial information that is required for machine learning algorithms. Therefore, it is wise to make use of patient's CT data stored as DICOM (Digital Imaging and Communication in Medicine) file. DICOM is the standard to store and communicate datasets in medical science. One can have better access to CT scan images in uncompressed version and get access to important information regarding patient's demographics from DICOM [3]. Again, with all precautions about data format, an inevitable element about CT images still exist in the form of unwanted additional information of surrounding body organs and external noises due to the different scanning devices used. Such irregularities have been recognized as a possible source of error for tissue classification [4]. To overcome these challenges and to make lung cancer detection easy, the entire work is divided into two subgroups as: image preprocessing and machine learning.

Set of sequential methods implemented for data normalization and segmentation composes image preprocessing. Input data normalization is the first step aimed to bring images from various sources to a regular standard by resampling pixel data contained in DICOM file of patients. It is followed by segmentation required to depict the lung abnormal regions and boundaries of the lung from surrounding thoracic tissues. Threshold-based segmentation is the most basic, well understood, and effective

S. Bhattacharyya et al. (eds.), *Advanced Computational and Communication Paradigms*, Lecture Notes in Electrical Engineering 475,
https://doi.org/10.1007/978-981-10-8240-5_37

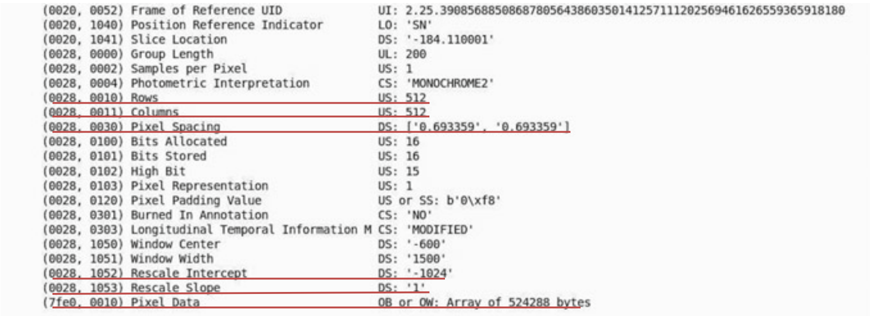
technique for obtaining segmentation from images with a well-defined contrast difference among the regions. This method segments the image by creating binary partitions that are based on image attenuation values and grouping together all the image elements to a region that satisfies the threshold interval. It is efficient as it takes only a few seconds to yield complete reproducible segmentation [5]. The coherency of the segmented image is not guaranteed and still may have holes, extraneous pixels, and noises. Region-growing segmentation is slightly advanced and effective method which serves as an efficient tool for extracting homogeneous regions with more precise lung segmentation results without false values. In this method, one pixel is compared with its neighboring pixel and if the predefined criterion (homogeneity) is met, then the pixel is said to belong to the same class as one or more of its neighbor. This method is useful for their efficiency and robustness in dealing with attenuation variations by reinforcing spatial neighborhood information and region criteria. Therefore, it is considered for segmentation and extracting pulmonary lung lesions in this paper. Other methods include shape-based segmentation which is modeled to give high accuracy, but are computationally expensive and performance highly depends on the feature set and training data. Two more segmentation methods being considered for clinical practice are neighboring anatomy-guided methods and machine learning-based methods. Both works well for specific conditions like when attenuation-matrices fails and to classify ground-glass opacity nodules, aided they require high computational power. To give a detailed idea of all methods is beyond the scope of this paper, but one can refer literature survey for more information on them [5, 6]. Finally, all preprocessed data is recorded along with a label set which contains labels for each input patient having cancer or not. Convolutional Neural Network (CNN) based on deep learning technique is used for training patients dataset. CNN having the advantage of both classification and prediction of input serves a robust and reliable algorithm to train against large datasets as CNN has N number of parameters to be manipulated for accurate training.

2 Materials and Methods

2.1 Image Preprocessing

2.1.1 DICOM and Hounsfield Units Conversion of Input Pixel Data

DICOM basically is a software standardization, which is used for image diagnosis and communication of image data using standard network protocol. The DICOM dataset of patients who undergo medical CT scanning serves as abundant amount of information related to his/her body organ which can effectively be utilized for image processing for diagnosis purpose. The biggest advantage to use this standard is the raw pixel image data along with other metadata about the patients and pixel values is present which further helps reduce a lot of computational complexity. Figure 1 shows the file view of a DICOM slice of a patient’s lung. The crucial information about pixel dimension and Hounsfield conversion factors, that is, rescale slope and rescale intercept are some important attributes of DICOM.



```
(0020, 0052) Frame of Reference UID              UI: 2.25.39085688508687805643860350141257111202569461626559365918180
(0020, 1040) Position Reference Indicator        LO: 'SN'
(0020, 1041) Slice Location                      DS: '-184.110001'
(0028, 0000) Group Length                        UL: 200
(0028, 0002) Samples per Pixel                   US: 1
(0028, 0004) Photometric Interpretation          CS: 'MONOCHROME2'
(0028, 0010) Rows                                US: 512
(0028, 0011) Columns                             US: 512
(0028, 0030) Pixel Spacing                       DS: ['0.693359', '0.693359']
(0028, 0100) Bits Allocated                      US: 16
(0028, 0101) Bits Stored                         US: 16
(0028, 0102) High Bit                            US: 15
(0028, 0103) Pixel Representation                US: 1
(0028, 0120) Pixel Padding Value                 US or SS: b'0\xf8'
(0028, 0301) Burned In Annotation                CS: 'NO'
(0028, 0303) Longitudinal Temporal Information M CS: 'MODIFIED'
(0028, 1050) Window Center                       DS: '-600'
(0028, 1051) Window Width                        DS: '1500'
(0028, 1052) Rescale Intercept                   DS: '-1024'
(0028, 1053) Rescale Slope                       DS: '1'
(7fe0, 0010) Pixel Data                          OB or OW: Array of 524288 bytes
```

Fig. 1. A DICOM file representing a slice of patient's lungs

Attenuation coefficient of material describes the fraction of incident X-ray beam being scattered or absorbed by the material as the beam penetrates inside. This helps in determining the material thickness and the material itself [7].

The Hounsfield Unit (HU) scale is a linear transformation of the original linear attenuation coefficient measurement into one in which the radio density of distilled water at Standard Temperature and Pressure (STP) is defined as zero Hounsfield Units (HU), while the radio density of air at STP is defined as 1000 HU. This is a universally accepted standard for all the CT scan tomography scanners which determines the effect of attenuation coefficient on the input intensity [8]. In short, it is a scale for identifying material whose values are recorded as attenuation coefficient. The relation between HU values and the attenuation coefficient is a linear equation described as below

$$y = m * x + c \tag{1}$$

where y is Hounsfield Unit value of pixel attenuation coefficient x, m, and c represents rescale slope and rescale intercept shown in Fig. 1, respectively.

Thus, conversion of input attenuation coefficients to HU values is evident to understand the nature of what body composition is present in the scanned data and also to know how scattered is the region of interest, lung nodules in this case. The theoretically prepared Hounsfield units scales which apply to medical-grade CT scans is used as reference to plot the histogram plot for the HU values of input scan dataset of a patient.

2.1.2 Resampling of HU-Converted Values

HU-converted input values are to be resampled as because of difference in scanning method and devices used, input data of different patients show different pixel spacing values along the 3 axes. Thus, bringing these values to a uniform value assures that the dataset of various patients are uniform structures free from errors generated because of scanning methods. When resampling is done the density of pixel in a patient's data changes, thus bringing a change to attenuation values and corresponding HU values. Now, by choosing a standard new spacing value we can easily get uniformity within our datasets. In general, the standard value of new spacing along the 3 axes is chosen as

(1, 1, 1) and existing values are converted to this, the algorithm for n patients data is discussed as follows.

```
Begin
Declare two list variables Slices, Images and initialize them with list of DICOM slices of patients and their pixel data respectively
Declare 1D arrays pixel_spacing, new_pixel_spacing, resample_factor, new_shape and resize_factor
Initialize new_pixel_spacing with [1, 1, 1]
for (i = 0, slice = Slices[i], image = Images[i] to i = n) do
pixel_spacing = Pixel Spacing of slice
resample_factor = pixel_spacing/new_pixel_spacing
new_shape = Shape of image
resize_factor = new_pixel_spacing/Shape of Image
Resample all pixel data in image by resizing them with resize_factor
Until termination condition is false
End
```

2.1.3 Segmentation of Resampled Pixel Data

The pixel data array which gets constructed after scanning of patients and the images that we plot from such data are seen to have many unwanted portion of pixel values which is of no use for diagnosis, that is, those are because of the images being square and scanners being oval, therefore this unwanted portions like air, bones, hard tissue, etc., are of no interest in the diagnosis of patient lungs. Therefore, segmenting necessary lung portions reduces the complexity of our algorithm and makes us have the region of interest necessary for analysis. Segmentation comprises of thresholding, connected component analysis [9], and removal of unwanted region. Figure 2 illustrates the segmented pulmonary nodules of patient's lungs. Algorithm for region-based segmentation [10] is discussed as follows.

```
Begin
Declare and Initialize seed variable SEED with position of seed (x, y)
Declare counter RCOUNT, stack PX, BP to keep track of current region, store pixels to grow and boundary pixels of region grown respectively
Declare array REGION and CP to store labels of grown region and 8-neighbours of CP respectively
R, C = size of the image
SEED = (R/2, C/2)
RCOUNT = 1, i = 1, j = 1 and PX(i) = SEED
Set threshold value THR as -320
while(PG not empty) do
CP = PG(i) and i = i - 1
```

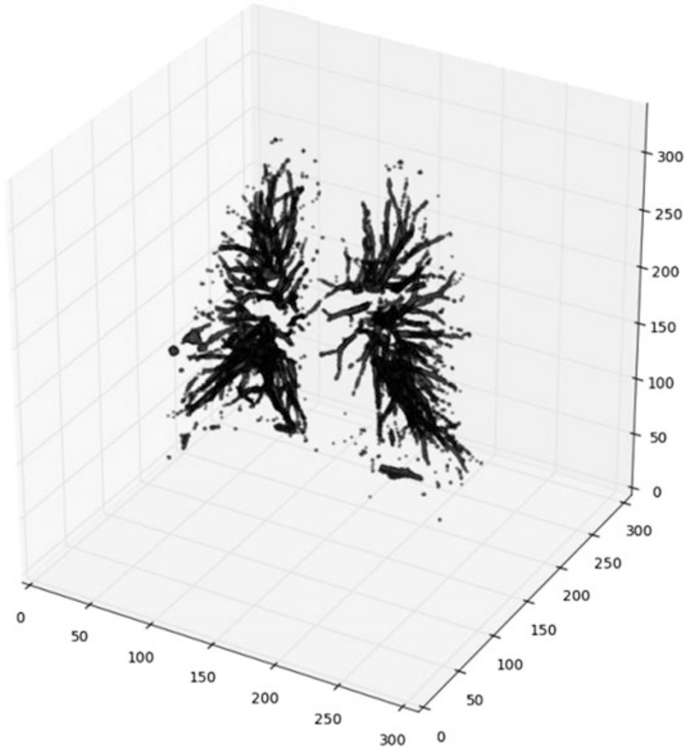


Fig. 2. Segmented pulmonary lung nodules

```
11 for (8-nb of CP, k = 1 to 8) do
12 If (REGION (CP(k)) is not labeled) then
13 Calculate the distance DIST of CP(k) from SEED
14 If (DIST > THR) then
15 REGION (CP(k)) = RCOUNT and i = i + 1
16 Else j = j + 1 and BP(j) = CP(k)
17 End for
18 End While
19 While (BP not empty) do
20 SEED = BP(j), j = j - 1, RCOUNT = RCOUNT - 1
21 i = 1 and PX(i) = SEED
22 Go to step 9
23 End while
24 End
```

2.1.4 Resizing of Images

The input dataset of each patient contains slices of resolution 512 × 512 and the number of slices for each one of them varies from 100 to 400 because of different calibration of the instruments from which the scan was taken, this nonuniformity is highly undesirable as the construction of neural network to perform operation of feature extraction on the input data depends on the size of the matrix passed as the input. Otherwise, every patient will require constructing a new neural network, which contradicts the training of machine. Thus, we conclude that this nonuniformity is highly undesirable and can be handled by downsampling or upsampling the input data to a certain fixed value across the depth.

2.1.5 Standardization

The standardization or feature scaling is important to step up the processing of machine learning algorithm and enhance the time taken for classification of datasets. It basically helps us reach values which bring easy path of convergence of weight values in training the machine using neural networks, thus reducing the time to predict and classify the data. This can be easily achieved by using following equations

$$\rho = \omega - \mu \tag{2}$$

where ρ is the result of subtraction of segmented pixel values ω from their mean μ.

$$\rho = \rho \div \sigma \tag{3}$$

where σ is the standard deviation of segmented pixel values ω.

3 Machine Learning

To train a Convolutional Neural Network (CNN) for extracting features from the input images by implementing deep learning technique is the basis of image classification and prediction of data. The design of CNN is illustrated in Fig. 3 for a better understanding.

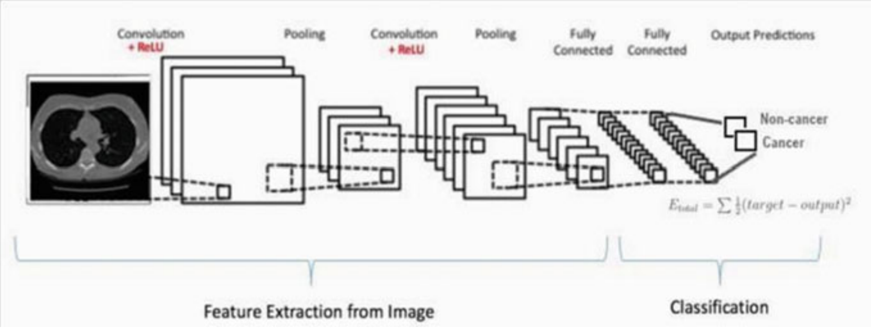


Fig. 3. Design of convolutional neural network [11]

The features extracted from convolutional part of CNN are used for prediction by passing them through a fully connected network and finally giving output with respective number of classes in the output layer of the network; two in this case. The reduction of loss function by changing weights using backpropagation algorithm supported with Adam optimizer for better convergence of weights is the real goal of training the network with correct weight values. The detailed understanding of CNN is necessary to bring it in use, and the study of CNN is very much limited in this paper as because, N number of parameters involved for tuning the constructed network to give accurate results. One can see the literature in references for further in-depth knowledge on CNN. The activation function is one key parameter of CNN which gives network nonlinearity aspect. The choice of activation function is independent, but results obtained by using Rectified Linear Unit (ReLU) are more promising. ReLU overcomes the limitation of traditional sigmoid and hyperbolic function, that is, the point of saturation which stops further change in weight values. Figure 4 shows ReLU curve.

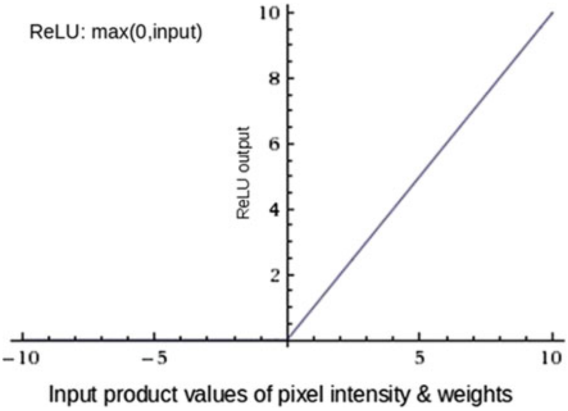


Fig. 4. Rectified linear unit curve

4 Experimental Results and Discussions

The proposed machine learning algorithm was tested on available dataset of 20 patients having cancer and no-cancer. The experimental results from training depict a gradual decrease in the loss function as shown in Fig. 5.

The performance evaluation of the proposed method resulted in detection accuracy of approximately 65%. The occurrence of lower accuracy is mainly due to the use of limited hidden layers in CNN computation. These layers are responsible for deep feature extraction depending upon the number of layers used, thus limiting to only 2 hidden layers might have failed to detect deep features accounting to cancer in the lung tissues. Also to mention that region-growing algorithm is good at detecting nodules in the lung region, but accurate extraction of nodules attached to lung walls, detection of

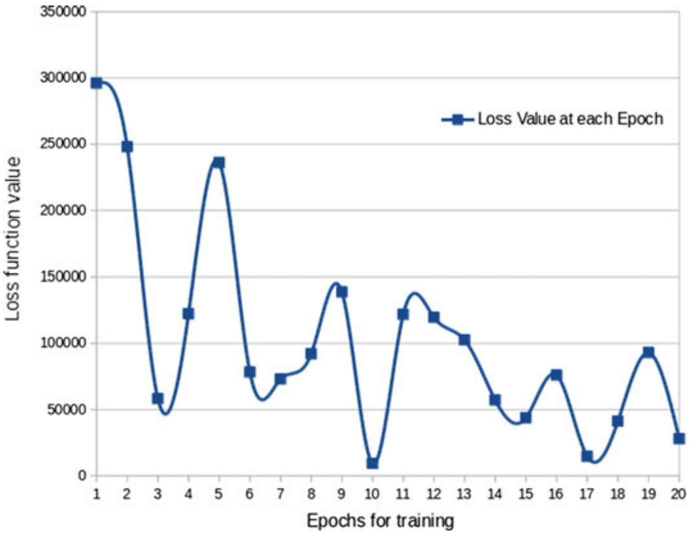


Fig. 5. Showing loss curve with respect epoch

ground-glass opacity nodules and to distinguish closely attached nodules and vessels are limitations of region-growing algorithm.

5 Future Scope

The image preprocessing can be enhanced to classify nodules better by detecting ground glass opacity nodules (which are deformed in shape but malignant), nodules attached to lung walls, anisotropic nodules, and also the ones which are closely attached with blood vessels. To attain this mean shift analysis based on the dot filter method of segmentation and Hessian matrix can be implemented.

The loss function values can result in a smooth curve with a gradual decrease in loss value by increasing the epoch size and once again by increasing the number of hidden layers in convolutional step more hidden features can be extracted from the input for better classification of patient's data.

6 Conclusion

Image classification and feature extraction from images to make accurate prediction is challenging and beneficial at the same time, it gives the ability to define objects, predict variation in images accurately at very deep dimension and also by leveraging the advanced methods of Neural Network (CNN), it is now possible to build an AI which can help reduce the burden on radiologist, analyst, hospitals, and clinical practitioners. Image networks is a special area of study for training machines to learn image data to classify and predict correct results, by bringing in advance additions to existing networks and also by making use of medically accurate segmentation methods it will definitely be possible to attain an accuracy of 95% and above for a wide variety of patients dataset.

Acknowledgments. We would like to thank Big Data team of System Level Solutions (SLS) Pvt. Ltd. Gujarat, India (www.slscorp.com), for their fruitful support, guidance, and mentorship towards the completion of this problem. We also would like to heartily thank SLS for giving us a noble opportunity to work on such a good problem statement.

We would like to thank Mr. Tejas Vaghela, General Manager, System Level Solutions Pvt. Ltd. Gujarat, India for his throughout support and guidance in this project.

We would like to thank Ms. Srijana Pradhan, Assistant Professor-I, CSE Department, Sikkim Manipal Institute of Technology for fruitful discussion.

References

1. Statistical Data of Lung Cancer. http://www.wcrf.org/int/cancer-facts-figures/data-specific-cancers/lung-cancer-statistics. Accessed 5 May 2017
2. Miah Md. BA, Yousuf Md. A (2015) Detection of lung cancer from CT image using image processing and neural network. In: ICEEICT, May 2015
3. The DICOM Standard (2017) Medical imaging and technology alliance—a division of NEMA
4. Horwood AC, Hogan SJ, Goddard PR, Rossiter J (2001) Image normalization, a basic requirement for computer-based automatic diagnostic applications. University of Bristol, UK
5. Mansoor A, Bagci U, Foster B, Xu Z, Papadakis GZ, Folio LR, Udupa FK, Mollura DF (2015) Segmentation and image analysis of abnormal lungs at CT: current approaches, challenges, and future trends. Radio Graph 35:1056–1076
6. Mesanovic N, Grgic M, Huseinagic H, Males M, Skejic E, Smajlovic M (2011) Automatic CT image segmentation of the lungs with region growing algorithm
7. Transmitted Intensity and Linear Attenuation Coefficient. https://www.nde-ed.org/EducationResources/CommunityCollege/Radiography. Accessed 13 May 2017
8. Hounsfield Scale. https://en.wikipedia.org/wiki/Hounsfield_scale. Accessed 15 May 2017
9. Panpaliya N, Tadas N, Bobade S, Aglawe R, Gudadhe A (2015) A survey on early detection and prediction of lung cancer. Int J Comput Sci Mobile Comput IJCSMC 4(1):175–184
10. Verma OP, Hanmandlu M, Susan S, Kulkarni M, Jain PK (2011) A simple single seeded region growing algorithm for color image segmentation using adaptive thresholding. IEEE
11. Karpathy A, Johnson J, Li F-F (2016) Convolution neural networks for visual recognition. Stanford University