(IJRST) 2019, Vol. No. 9, Issue No. III, Jul-Sep **e-ISSN: 2249-0604, p-ISSN: 2454-180X**

COMPARATIVE ANALYSIS OF LEARNING ALGORITHMS FOR LUNG CANCER IDENTIFICATION[1]

Yash Jaiswal

Undergraduate student, Galgotias University, Greater Noida, India

Received: 21 May 2019; Accepted: 08 August 2019; Published: 27 August 2019

ABSTRACT

Lung Cancer detection making use of medical imaging is still a challenging task for radiologist. The objective of this research is to classify the types of lung tumours for extracted and selected features using learning algorithms. In this paper, an experimental study is conducted on 100 cases of lung cancer to evaluate the performance of learning classifiers (DNN, SVM, Random Forest, Decision Tree, Naïve Bayes) with different medical Imaging (DICOM) features to identify the two types of Lung cancer (Benign and Malignant). The proposed system means to robotize the whole strategy of determination via naturally identifying the tumor, estimating the necessary qualities, for example, breadth, edge, region, centroid, roundness, spaces, and calcification. The trial is led into two stages: In the principal stage, distinguish the most critical component utilized in lung cancer analysis by CT scan and perform the mapping to computer related format. In the second phase, feature selection and extraction is performed to machine learning algorithms. To evaluate the performance of classifiers in term of classification accuracy and improving the false positive rate, every stage of evolution is divided into four different phases: single phase module, single slice testing, series testing and testing of learning algorithms. Experimental results show significant improvement in false positive rate up to 30% for both Benign and Malignant. Whereas, Deep Neural Network (DNN) demonstrate high values in term of classification accuracy in comparison with other classifiers. The proposed methodology for lung cancer detection system having a potential to reduce the time and cost of diagnosis procedure and use for early detection of lung cancer.

1. INTRODUCTION

Cancer is characterized as an illness that features anomalous cell development pursued by flighty cell propagation. It is one of the significant reasons for death internationally. Typical cell physiology comprises cell development, division, and apoptosis in an arranged way. In the event that this physiological procedure gains out of power, cells develop too quickly with no structure and form into an irregularity which is known as a tumor. Tumors can be ordered into two gatherings: benevolent and harmful. While a kind tumor is confined to the site of development and doesn't develop that quickly, a harmful tumor basically comprises of rapidly developing dangerous cells that can spread past the first site, attack neighboring tissues, and spread to different pieces of the body in a procedure called as metastasis1. Lung cancer growth screening is incredibly valuable since it broadens lives at a sensible expense and satisfactory degrees of hazard. A screening test ought to distinguish every single surviving disease while keeping away from superfluous workups2. Be that as it may, in this lies the issue: radiologists can't reliably and precisely analyze lung cancer growth when taking a gander at a CT scan; Interpretation of medicinal pictures is troublesome and tedious. The execution of picture handling methods and AI can make this procedure substantially more proficient. As computerized advancements are fused in each part of our lives, they can likewise turn into a key piece of medicinal diagnosis3. Among men lung, Cancer growth

[1] *How to cite the article:* Jaiswal Y., Comparative Analysis of Learning Algorithms for Lung Cancer Identification; *International Journal of Research in Science and Technology*, Jul-Sep 2019, Vol 9, Issue 3, 42-50

is the most widely recognized sort for both event and mortality. Among ladies, it has the third most noteworthy frequency and second most elevated death rate after bosom cancer growth. The death pace of lung malignancy is the most noteworthy among a wide range of tumors. It additionally has a littlest endurance rate after the finding among a wide range of diseases, with a continuous increment in the number of passings consistently. Out of around 220,000 instances of dangerous cancer growth consistently, very nearly 160,000 patients bite the dust. Endurance from lung malignancy is straightforwardly identified with its development at its identification time, so the odds of endurance increment if the disease is distinguished in the beginning times. The principal supporter of lung malignancy is smoking4. An expected 85 percent of lung disease cases in guys and 75 percent in females are brought about by cigarette smoking3. Different causes incorporate radon gas, asbestos, air contamination, hereditary qualities, and so on. The populace section well on the way to create lung disease is individuals matured more than 50 who have a background marked by smoking. A Computerized Tomography (CT) check is a mix of a set or a progression of X-beam pictures. These pictures are taken from various edges at various times. Cross-sectional pictures of cuts at various occasions are delivered utilizing PC handling or cut these cuts might be of the bones, veins and delicate tissues. It delivers a 3D picture of the human body. For recognition and determination of lung cancer growth, CT examines are said to be more compelling than plain chest X-beams. Dissimilar to customary X beams, which just feature thick body parts, for example, bones, CT additionally gives nitty-gritty perspectives on the body's delicate tissues, including veins, muscle tissue, and organs, for example, the lungs. Additionally, while ordinary X beams give level 2D pictures, CT pictures delineate cross-segments of the body5. The accompanying Figure 1 shows a regular CT filter picture.

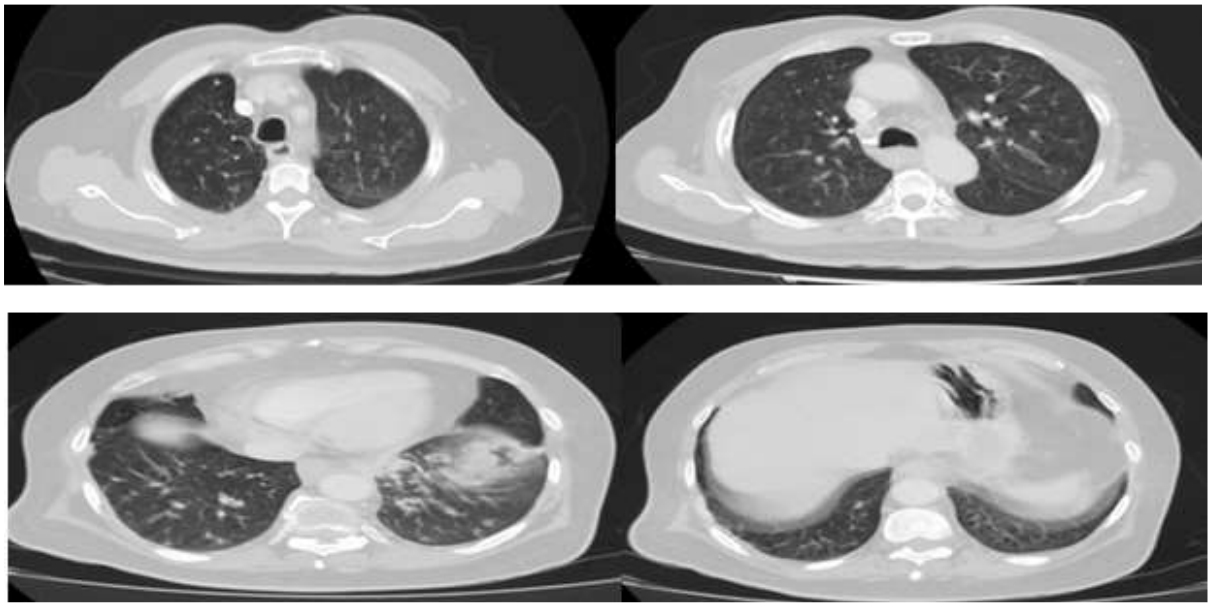


Figure 1. CT scans images taken at different time slice.

Lung cancer is second highest mortality rate and third highest incidence. Around 220,000 cases of malignant cancer are reported every year, out of which almost 160,000 patients die. To illustrate the problem that radiologists have, we are using a recent study conducted by the NIH. 53,454 smokers from ages 55 to 74 were all tested for lung cancer and all of them had CT Scans. Overall, 39.1% of the smokers had positive screens according to the radiologists. In other words, the radiologists thought that 20,901 patients had lung cancer. However, after a verification check, there was a horrendous false positive rate of 96.4%. This meant that out of the 20,901 patients diagnosed with lung cancer, only 752 patients actually had lung cancer. This approach intends to decrease the enormous false positive rate. The second reason to develop this system is to decrease the overall time of the diagnostic procedure. According to a research conducted by us in different hospitals of Pakistan we got to know that it takes about 4-5 hours for a radiologist to analyse a CT scan in order to give a verdict. Apart from that they view the CT scan on DICOM viewer and manually calculate the area and HU

values of the tumour. There is a chance of error while measuring the area manually by drawing over the tumour joining its edges. The objective of this research is to classify the lung tumours as benign, malignant for extraction and selected features using Learning algorithm. The comparative analysis among the learning algorithms will show reduce the enormous false positive rate by increasing the efficiency and accuracy of the diagnostic procedure. The proposed approach means to computerize the whole methodology of finding via naturally distinguishing the tumor, estimating the necessary qualities, for example, region, edge, width, centroid, roundness, spaces, and calcification and giving a forecast about the kind of tumor as appeared in Figure 2.

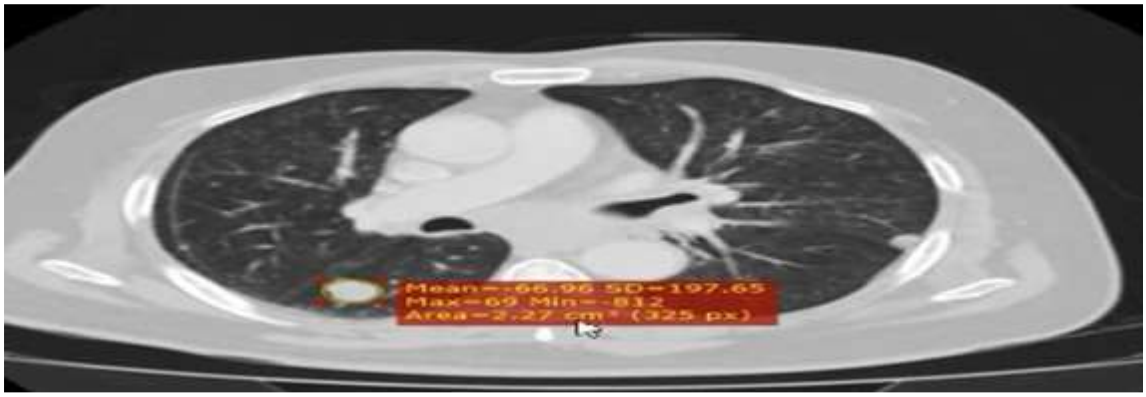


Figure 2. Manual tumour calculation by drawing a circle.

Rest of the paper is organized as follows: the consequent section provide the literature review. In section 3, design phases for developing experiments are defined. Different selected and extracted features are defined for lung cancer identification in section 4 and 5 respectively. Experimental results are presented in section 6. Finally, the conclusions are drawn in section 7.

2. LITERATURE REVIEW

Nowadays researchers are working on the development of a diagnose system that is based on sputum colour images. Numerous algorithms have been suggested for medical imaging. In4, Multiple image processing methods are used to detect lung cancer with the aid of computerized tornography scan images. This idea was proposed by Vijay A. Gajdhane et.al.in3. This procedure involved three processes majorly, Pre-processing, Feature Extraction and Classification. In the first step median filter was used to remove the noise, and with the aid of gabor filter image is segmented and enhanced. The features extracted were eccentricity, perimeter and area. In the last step, for classification SVM is used. In6 a SVM based classifier were proposed the help of image classification technique based on features. Lung nodules in LDCT slides are further classified into four groups that is, vascularized, well circumscribed pleural-tail and juxta-pleural. The suggested method based on Support Vector Machine (SVM) was trained by using C-SVC a polynomial kernel. Another way of using support vector based method was proposed by Hiram Madero Orczco et al7. He used the gray level concurrence matrix and eight texture features which were mined out from a histogram to identify and classify images as without nodules or with it. During process segmentation stage wasn't considered. Tiwari8 and Sharma9 were analysed lung CT images and proposed automatic CAD system for the detection of lung cancer. Sundararajan et al10 were focused on lung disorders and used textural features for disjoint sections of lung to propose a support vector machine that aided recognition of pneumoconiosis. Chaudhary et al11 used enhancement and segmentation techniques to obtain more precise results. Hashemi et.al. (2013) proposed a region growing segmentation method to improve lung cancer diagnosis. Schilham et al12 contributed in detecting lung nodules of different size by proposing the use of k-nearest neighbour (k-NN) classifier. Pereira et al13 were detected lung nodule using a sliding band filter based on the convergence of radial gradients. Vijai Anand et al recognized tumor as benign or malignant by a proposing a system that predicted lung tumour from CT images with the help of image processing techniques with neural network classification14. Lee et al proposed detection of all the nodules in the lung and recorded a low false rate using a random forest based classifier15. JIA Tong el al16 were identified the pulmonary nodule from CT images by an automatic Computer-Aided Detection (CAD) scheme. Mao et al has evaluated the nodule enhancing capabilities of fragmentary window filtering17.

3. DESIGN PHASE

To evaluate the performance of learning classifiers using selected features, the proposed design is divided into two different phases:

1. Feature extraction/selection, and

2. Making use of Machine learning classifiers and/or deep learning algorithm.

In the first phase, personnel from the two different areas of expertise were involved as radiologist and lung researchers. Radiologist: Provide help to understand the relative information of the anatomy, physiology of the lung functionality and the symptom of the cancers and their types derived from the available CT scans. The personnel from engineering discipline analysed the transformation of the parameters related to lung cancers (disease and their types) in to computer understandable parameters, which are used to drive the further stages for the development of tool. The knowledge/understanding of the following aspects was focused in this phase:

1. Identifying the most significant feature used in lung cancer analysis by CT scan,
2. Comprehensive literature review has been done to answer the following possible questions:

- Is there any specific tool available to retrieve the relative feature of the specified disease? How we can relate the parameters used in CT scan report analysis in to computer aided form and relate the change actually occurring in the physical test report accurately mapping in to computer related output?
- Identifying the best possible source for transferring the physical CT scan images into required computer format.

The second phase of this work processed the related information that was transformed from physical results of patients in to computer based learning algorithms. In Machine learning methods one of the significant and initial steps was to identify and select the appropriate feature for machine learning classifiers.

3.1 Feature Selection

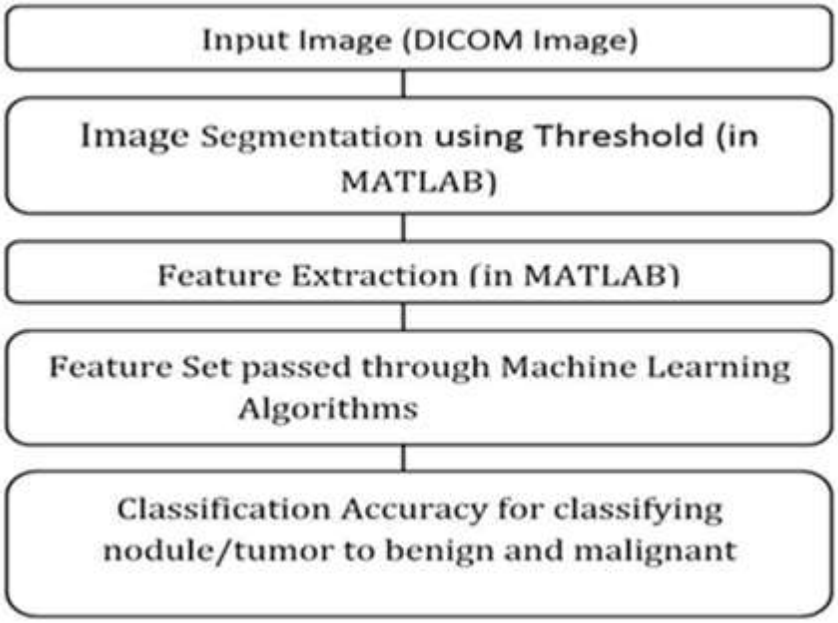
In general, the machine learning algorithm makes use of spectral features to classify the pattern of the different objects but it varies from application to applications. In this project one of the significant research oriented task was to identify the best or fairly relative feature which transforms the actual information of the CT scan result in to computer based outcome. Machine learning is used to achieve this goal and the selection of the feature is done from available feature stream and some are derived as new features using in-depth understanding of the physiology of the lung with and without disease.

3.2 Selection of Learning Classifiers

In this step, the machine learning tool used the derived features from the previous step to select the most appropriate or accurate classifier based on recognition percentage value. These results are used to develop the combination of different classifiers with derived features to select the most accurate classifiers in particular situation. The result of the classifier helped to identify lung cancer with appropriate features and learning classifiers. The following steps were used to implement the design phases:

1. The CT-Scan images of lung cancer patients along with their complete diagnostic summary were acquired
2. The DICOM images are fed into the system. The system analysis each slice of the image set.
3. The tumour area is highlighted and its features such as centroid, diameter, area, smoothness and indents etc. were extracted using image processing techniques.
4. The feature set is then passed to Machine learning algorithms for the identification

The procedure



4. FEATURE SELECTION

4.1 Lung Cancer Symptoms

The signs and symptoms of lung cancer are coughing, Haemoptysis, weight loss, wheezing or shortness of breath, fever, and fatigue. The side effects because of a malignancy mass pushing on adjoining structures: chest torment, bone agony, predominant vena cava deterrent, and trouble gulping. Indications that propose the nearness of metastatic ailment incorporate weight reduction, bone torment, and neurological side effects (cerebral pains, swooning, seizures, or shortcoming). Regular locales of spread incorporate the mind, bone, adrenal organs, inverse lung, liver, pericardium, and kidneys.

4.2 Lung Nodule Classification

There are two major classifications of a tumour. It can either be a benign or a malignant tumour Benign and Malignant. Basic aim of the presented software design is to differentiate between the two types of tumour. A benign tumour is not cancerous where as a malignant tumour is cancerous. The treatment of cancer depends on the type of tumour detected. The following characteristics of both tumour helped in feature selection.

4.3 Benign

- It is not cancerous,
- It is localized in a region and doesn't invade in other tissues,
- Small in size that is less than 2 cm (<2cm) and size doesn't change for 2 years,
- It is rounded in shape with smooth edges, and
- It has calcium deposits in it and the HU value is near to bone.

4.4 Malignant

- It is cancerous,
- It invades in surrounding tissues and may spread in the body,
- Size is more than 2.5cm (>2.5 cm),
- It has irregular and speculated edges, and
- It has no calcium deposits and HU values correspond to that of fluids and soft tissue.

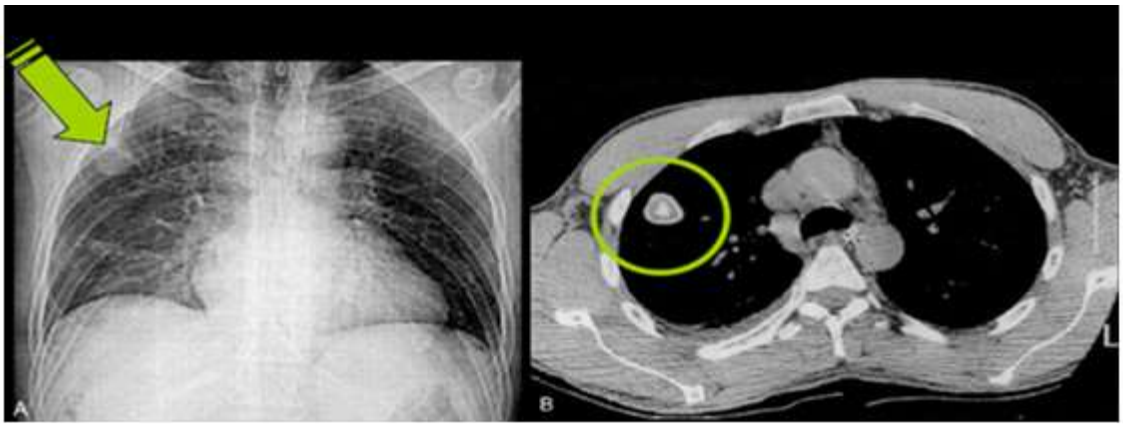


Figure 4. Benign pulmonary nodule with smooth edge and central calcification.

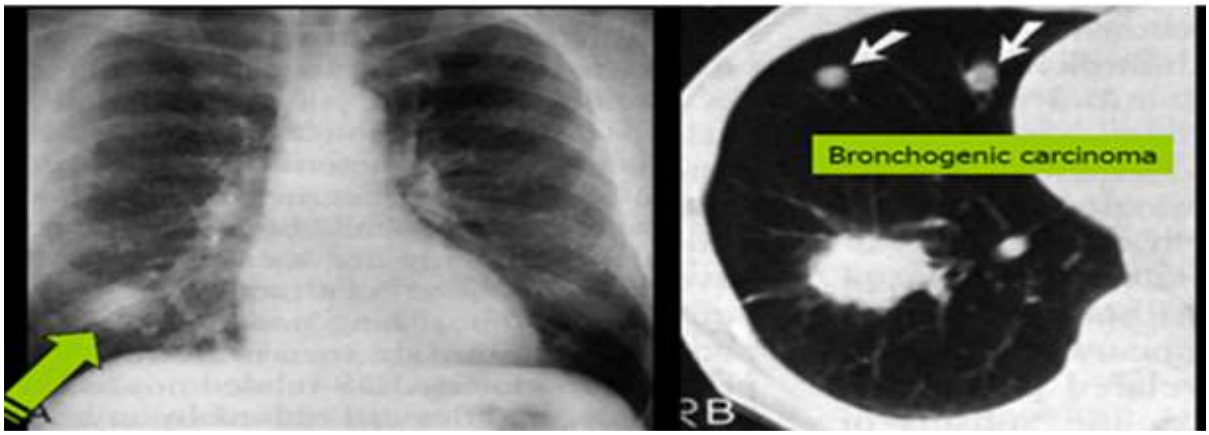


Figure 5. Malignant pulmonary nodule with speculated edge and ipsilateral deposits.

4.5 Hounsfield Unit

HU value, named after Godfrey Newbold Hounsfield, is a standard quantity used in medical world for CT scanning to map the CT numbers to a value in standardized form. It describes the radio-density. It linearly transforms the original linear authentication coefficient into a specific value. Zero Hounsfield Unit (HU) corresponds to the radio-density of distilled water at STP whereas -1000 HU correspond the radio-density of air at STP.

The HU values are calculated on the basis of the following formula18.

$$HU = 1000 \times \frac{\mu - \mu_{water}}{\mu_{water} - \mu_{air}}$$

4.6 Selected Features

The following features were selected on the basis of the above characteristics, Area, Perimeter, Diameter, Centroid, Roundness, Smoothness, Indentations and Calcification.

5. FEATURE EXTRACTION

5.1 Data-Set

We acquired the data set from different hospitals in Pakistan and a large part of the data was downloaded from the internet. We got about sixty cases from the internet but the data was quite old that was of 1990's.. Nine cases were provided by Karachi Institute of

Radiotherapy and Nuclear Medicine (KIRAN) Hospital. The data was provided with their diagnostic reports. However, the data downloaded from internet had no reports but was verified by the radiologist at Aga Khan University and Hospital (AKUH). The images are in the Digital Imaging and Communication of Medicine (DICOM) format. Initially the algorithm for images processing were developed using the downloaded dataset but was later on modified according to the recent data of 2017 acquired from the hospitals. The dataset is generated using 16 slice machine that is currently being used in many hospitals in Pakistan. It included many variations in cases. Some were adenocarcinoma, squamous cell carcinoma and large cell carcinoma.

5.2 Input DICOM Image

DICOM is a standardized set of rules that is being used worldwide, to define the storage, exchange and transfer of medical images. All different medical imaging devices like X-rays, MRI and CT etc. use the same standard. An image in DICOM format is useful as it not only stores the pixel values but also data set with some additional attributes. It may contain critical information like a patient's identity, diagnostic history and related information that may help in diagnosis. The DICOM images can be viewed using many different viewers. The one that is used in this research is Radiant DICOM Viewer, for analysis of dataset, before developing the algorithm for image processing.

5.3 HU Module

Once the lung is masked then the image returned from dicomread() is passed to HU module along with the information retrieved from the dicominfo()19. This function simply maps DICOM information which is the HU units on the corresponding intercept of the image which are obtained from the following:

```
info = dicominfo (file name)
intercept = info. Res
caleIntercept;
slope = info. Rescale Slope;
New pixel value is calculated as:
new_pixel = p'*s + i;
```

A new image is formed in which each pixel value corresponding to a HU value is mapped onto the intercepted values.

5.4 Lung Module

This module performs the basic morphological operations including dilation and erosion. This model is the core of the detection algorithm which detects the abnormality in the lung without filtering some other veins and arteries that may be confused to be a tumour. First the de-noising of the images is performed by a median filter, removing all background textures, nerves and lining of the lung which is irrelevant in detection. This de-nosing provides filtering mechanism where some of the nerves and arteries. Next the image is thresholder w.r.t to other pixels in a way that a new pixel is picked up and is compared to other pixels. If the value of pixel is less than the picked up pixel then it is set to 0 otherwise 1. This way a binarized image is formed. After binarzing there are some grains left inside the binarized, to remove them a method of morphological operation is used named as closing which says that dilation should be followed by erosion to fill out incomplete hole inside the figure. This resulted in solid pixelated figure without holes. Since the nerve lining in the lung are relatively small as compared to the size of the tumour they are filtered out by built in function in MATLAB bwareaopen (). The purpose of the function is to filter out the figure which has an area of pixel smaller than the value specified in the parameter of the function. These results in a final picture which gives ROI that are tumours, if they are detected. The boundaries of the figure are defined by drawing a line over the edges of the pixel that result in a clear view of the detected tumour. The feature of the detected tumour are then computed using built-in region prop () function that returns area, perimeter, centroid and diameter.

5.5 Location of Tumour

The complete dimensions of the lung image are taken and it is broken in to seven columns and five rows which make a total of 35 cells. Each individual cell is labelled against the location of the lung. Labelling is based on three major subheads per cell.

This is performed on each slice i.e. 16 slices so we get a combination of 35*16 different locations of lung. Each slice is mapped onto a specific region of lung. This helps in specifying the location of tumour in the lung. Whenever a defected region is detected the slice count along with coordinates of the tumour are passed to this module.

5.6 Sphercity

It is categorized in two parts
Indentations and

Roundness

5.7 Roundness

This function is passed a cropped image of the tumour and not the entire lung. It sets the contrast to a specific value to enhance the visibility of the image. It then creates a circle around the tumour and gives a probability that how much the figure is close to a round. If it is close to circle it indicates that it may be a benign otherwise a malignant.

5.8 Indentations

Indentations are used to see whether a tumour has got speculated edges or not. If the number of indentations is greater this means that there is speculation in the tumour. This is a characteristic of malignant tumour. This module joins the edges of the tumour by a line in a way that it forms a new boundary as shown in the figure. After that it calculates the number of indents from the tumour boundary to the line. Depending upon the number of indents it differentiates between malignant and benign.

5.9 Calcification

Calcification may be defined as the accumulation of calcium in a body tissue. It is the calcium deposit in a tumor. A benign tumor has calcified part i.e. solid calcium deposit in it whereas a malignant tumor doesn't have calcification in it. In this module the RGB value are manipulated to detect calcification. The RGB values are set in a way that it highlights the bony structure. It the value matches a threshold value it detects the calcified part just as done in the lung module. The outlined calcified part is passed to regionprop () to determine the area. This area is compared with the area of tumor giving a percentage of how much part of the tumor is calcified20.

6. EXPERIMENTAL RESULTS

To evaluate the performance of learning classifiers, the testing of every stage of evolution is divided into four different phases:

6.1 Single Phase Module

The first phases were to test each individual module that was developed for image processing. Seven different modules have been implemented for image processing. Each module was tested to check whether it is performing as expected or not. If the module didn't produce the expected result it was re-implemented in order to produce correct results,

6.2 Single Slice Testing

In this phase each slice from the series of image slices was tested individually so that we can differentiate that which slice produces the exact result as generated by the radiologist in their analysis. There is a chance of error when the radiologist calculates the area of a tumour since they measure the area manually by drawing a line. This testing formed the basis of the implementation of the compiled version of all the series,

6.3 Series Testing

This phase was the final testing of the image processing results produced. The feature that was extracted using the image processing techniques was compared with the actual verdict given by the radiologist. We were provided with the reports of each case so it was easier for us to compare the result and calculate percentage error between the results, and

6.4 Testing of Learning Algorithms

This is the final testing phase of the entire system. The results produces at this stage i.e. a probability provided by the system about the tumour being a benign (no-cancerous) or malignant (cancerous), were cross-checked with the results provided by the radiologist. If the malignant probability was more for a case of malignant that means that system is working correctly. We had no false negative rate in the system.

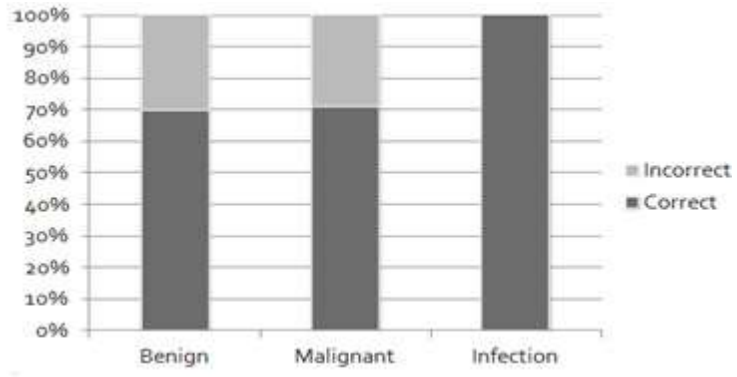


Figure 6. Comparatively analysis of boning, malignant and infection.

To implement the proposed methodology: the total of 100 cases of lung cancer for training and testing of the system. That data set was very limited which affects the accuracy of the system out of 100, 90 cases were used as training data and about 10 cases for testing. The following graph shown in Figure 6 represents that our false positive rate is not more than just 30% which is a great deal.

There is a false positive rate of 30% for both malignant and benign whereas the system never diagnosis an infection as tumour. Our system doesn't detect infections at the moment but it is efficient enough to differentiate between the characteristics of tumour and infection. It is free of false negative rate which is a great achievement at this stage.

The experiment has been performed using learning machine with five different algorithms namely deep neural network, Random forest, Naïve Bayes, Support vector machine and Decision tree.

Table 1. Comparative study of classification accuracy

Algorithm	Accuracy
Deep Neural Network	88.5%
Naïve Bayes	64.28%
Support Vector Machine	62.5%
Decision Tree	58%
Random Forest	65.4%

A comparative study shows in Table 1 that; the highest accuracy result is provided by the deep neural network. Second is Random forest. But there is always a trade-off between time and accuracy when it comes to machine learning algorithm. We need to optimize the system in such a way that the machine learning algorithm takes least time and provides greater accuracy. The following table shows that comparative analysis of the above mentioned algorithms.

7. CONCLUSION

The accuracy levels showed by this proposed system are considerable. The false positive rate comes out to be 30% which is far better than the doctor's 96.4% false positive rate in the study conducted by NIH. The accuracy of the proposed system is currently compromised due to very limited corpus but we can optimize the overall system by testing the software on large amount of data. The compromised accuracy would obviously be changed when the dataset grows. Authors are further working on improving image processing and machine learning techniques to increase the accuracy of the system and decrease the overall false positive rate. This system has potential to reduce the enormous false positive rate of lung cancer detection and facilitating radiologists in making their decision more accurately. The proposed system will also cut down the overall cost of the diagnosis procedure, getting biopsy proven reports of the patients along with the CT scans there is a chance that our system will decrease the probability of undergoing a biopsy and use for early detection of lung cancer which can save many lives from the expense of diagnosis, pain of treatment and even death.

Financial Support and Sponsorship: Nil

Conflict of interest: None