Early Detection of Lung Cancer using Sputum Cytology

Dharmesh A. Sarvaiya[1] Prof. Mehul Barot[2]

[1,2] Department of Computer Engineering

[1,2] L.D.R.P Institute of Technology & Research, KSV University, Gandhinagar.

Abstract— Lung cancer is acknowledged to be the fundamental driver of disease passing worldwide, and it is difficult to detect in its early stages because symptoms appear only in the advanced stages causing the mortality rate to be the highest among all other types of cancer. The early detection of cancer can be helpful in curing disease completely. This study paper summarizes various reviews and technical articles on Lung cancer detection using the data mining techniques to enhance the Lung cancer diagnosis and prognosis. The present work deals with the attempt to detect lung cancer at early stage based on the analysis of sputum color images The recognition of lung tumor from sputum images is a testing issue because of both the structure of the disease cells and the stained strategy which are utilized in the definition of the sputum units., This survey paper includes the survey of different techniques such as threshold classifier, a Bayesian classification and Hopfield Neural Network and Fuzzy C-mean.

Key words: Lung cancer detection, sputum images, threshold technique, Bayesian classification, Hopfield neural network.

I. INTRODUCTION

Lung cancer is the uncontrolled growth of abnormal cells that start off in one or both lungs; usually in the cells that line the air passages. The abnormal cells do not develop into healthy lung tissues; they divide rapidly and form tumours. As tumours become larger and more numerous, they undermine the lung's ability to provide the bloodstream with oxygen. Tumours that remain in one place and do not appear to spread are known as "benign tumours". Malignant tumours, the more dangerous ones, spread to other parts of the body either through the bloodstream or the lymphatic system. Metastasis refers to cancer spreading beyond its site of origin to other parts of the body. When cancer spreads it is much harder to treat successfully. Lung cancer symptoms consist of shortness of breath, wheezing, chest pain that does not get better, coughing accompanied with blood, difficulty in swallowing, and loss of weight and appetite [1].

Lung cancer can be broadly classified into two main types based on the cancer's appearance under a microscope: non-small cell lung cancer and small cell lung cancer. Non-small cell lung cancer (NSCLC) accounts for 80% of lung cancers, while small cell lung cancer accounts for the remaining 20%[1].Lung cancer occurs when a lung cell's gene mutation makes the cell unable to correct DNA damage and unable to commit suicide. Mutations can occur for a variety of reasons. Most lung cancers are the result of inhaling carcinogenic substances.

Most lung cancers are first diagnosed based on symptoms. Symptoms of lung cancer are not very specific and generally reflect damage to the lungs' ability to function normally There are many techniques to diagnose lung cancer, such as Chest Radiograph (*x*-ray), Computed Tomography (CT) , Magnetic Resonance Imaging (MRI scans) and Sputum Cytology. On the other hand, the vast majorities of these methods are costly and time intensive. Unlike mammography for breast cancer or colonoscopy for colon cancer, a widely accepted screening tool for early-stage lung cancer has not been available until recently. Regular chest X-rays are not reliable enough to find lung tumours in their earliest stages, when many doctors believe the tumours are at their smallest and most curable state.

Among all the previous techniques, only the manual sputum cytological examination has been utilized for the diagnosis of early lung cancer detection since 1930s [2]. Sputum cytology examines a sample of sputum under a microscope to determine whether abnormal cells are present. Sputum is not the same as saliva. Sputum is produced in the lungs and in the airways leading to the lungs. Sputum has some normal lung cells in it.

A sputum sample may be collected by [3]:

1) By a person coughing up mucus.
2) By breathing in a saltwater (saline) mist and then coughing.
3) During bronchoscopy, this uses a bronchoscope to look at the throat and airway.

After a lung cancer is suspected based on imaging, a sample of tissue is required to confirm the diagnosis and determine the type of cancer. Sputum cytology is the easiest way to do this, but its use is limited to those tumours that extend into the airways. In this paper, we focus on different techniques to detect lung cancer at early stages with use of sputum cytology.

The automatic of a sputum cell state is based on the analysis of its nucleus and cytoplasm. The sputum cells are characterized by uncertainty cells pattern that make the segmentation and detection of the cells very problematic, so it is difficult to segment the foregrounds from the image automatically and perfectly [4].

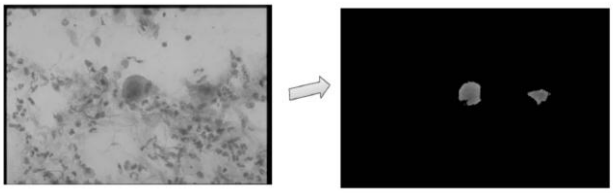

Fig. 1: Sputum Cell Detection

Sputum represents a highly specialized product of the respiratory tract. In patients with sputum production, one can glean information pertinent to the abnormality through

the intelligent examination of this material. Cytological examination of Papanicolaou stain of sputum is accepted as a useful diagnostic tool in carcinoma of lung.

II. RELATED WORK

For Early detection of Lung cancer, there are different approaches of data mining to detect nuclei of the sputum image and cluster those images.

In [5], researchers have proposed a Computer Aided Diagnosis (CAD) system for early detection of malignant lung cancer cells using digital images of stained sputum smears. Such an automated system would allow objective and unbiased assessment, as compared to human evaluation which might be corrupted by errors originating from inter-and intra-observer variability that characterizes human observation. Eventually, this system will be useful for handling large sputum image databases and relieving the pathologist from tedious and routine task. In this paper, they focus on the extraction and segmentation of sputum cells from background regions. The sputum images are stained according to the Papanicolaou standard staining method.

These images are stained with two types. Type1, blue dye images resulting in the dark-blue nucleus of all the cells present in the image and clear-blue cytoplasm. Type 2, red dye images resulting in the dark-blue nucleus of the small debris cells with their corresponding small clear-blue cytoplasm regions, and red sputum cell with dark-red nucleus and clear-red cytoplasm. Some of the sputum nuclei cells overlap due to the dispersion of the cytoplasm in the staining process [5].

In this paper they propose two methods for addressing this problem the first employed a threshold-based technique. The second method uses a Bayesian classification framework. The problem of extracting the nucleus and the cytoplasm is approached using a combination of robust mean shift segmentation and rule-based techniques.

A. Sputum Cell Detection [5]

The unit discovery points at the extraction of the cell area from the sputum picture. This is carried out by verifying whether a pixel in the sputum picture fits in with the sputum cell utilizing its shade data. The staining strategy, connected in the sputum specimen result, permits, to some level, the sputum cell to have a unique chromatic manifestation opposite the foundation. In any case, the way of the sputum shade pictures, which holds numerous trash units and the relative complexity around the cytoplasm and cores cells, implies that the extraction process for the cores and cytoplasm units is not a direct system.

1) Threshold Technique [5]

The threshold technique depends on the staining methods by which the image is organized and derived from the difference in the brightness level in RGB components of the sputum color images. Here The parameter Θ is determined by trial and error testing whereby the outcome of the segmentation is assessed visually.

For the image stained with blue dye, the following rule is used to extract sputum pixels:

If (B(x, y) < G(x, y) + Θ) then B(x, y) is a sputum

Else B(x, y) is non sputum.

For the image stained with red dye, where the red color is the most dominant color between the sputum cells and the background, we use the following sequence of rules:

If (B(x, y) < G(x, y) or (B(x, y) > R(x, y)) then B(x, y) is

Sputum else B(x, y) is non sputum.

The optimal value of the threshold was determined by analyzing the performance of the method across.

2) Bayesian Classification [5]

In this approach they address the cell detection problem using a probabilistic method based on the Bayesian classification.

In these methods, a pixel x is considered part of the sputum region if p (bg | x) < p (sp | x) where sp and bg refer to the sputum and the background respectively. Applying the Bayesian Rule and the concept of classification cost, this inequality can be brought to [5]:

$$\sigma = \frac{\mu sp}{\mu bg}\frac{p(bg)}{p(sp)} < \frac{p(x|sp)}{p(x|bg)} \quad (1.1)$$

where is μ_{sp} the loss weight incurred if the sputum class has been selected instead of the background and μ_{bg} is the loss weight incurred if the background class has been selected instead of the sputum, p (bg) and p (sp) are the probabilities of the background and the sputum classes respectively, and they are estimated from the total number of sputum and background pixels in the training set of images according to the following equations:

$$p(\text{sp}) = \frac{Tsp}{Tsp+Tbg} \quad (1.2)$$

$$p(\text{bg}) = \frac{Tbg}{Tsp+Tbg} \quad (1.3)$$

where Tsp and Tbg are the numbers of sputum and background color respectively.

A database of 100 images, collected from the Tokyo Center for lung cancer, was utilized in this study. The size of each image is 768 × 512 pixels and they were provided in the RGB space. They conducted a comprehensive set of experiments to study the outcome of the threshold algorithm for the detection and extraction of the cells into sputum cells and background. Furthermore, they analyzed the essence of color representation and color quantization on the sputum cell detection. Then they used the cell extraction techniques the Bayesian classifier.

The Bayesian classification achieved the best scores. It succeeded particularly in reducing the number of False Negative and improving the sensitivity. On the other hand, the specificity and accuracy are close to their counterparts in the threshold methods. Bayesian classification achieved sensitivity-89%, specificity-99%, accuracy98% [5].

In [6], the researchers have proposed two segmentation methods, Hopfield Neural Network (HNN) and a Fuzzy C-Mean (FCM) clustering algorithm, for segmenting sputum color images to detect the lung cancer in its early stages. The segmentation results will be used as a base for a Computer Aided Diagnosis (CAD) system for early detection of lung cancer which will improve the chances of survival for the patient. we applied a

thresholding technique as a pre-processing step in all images to extract the nuclei and cytoplasm regions, because most of the quantitative procedures are based on the nuclear feature. The HNN and FCM methods are designed to classify the image of *N* pixels among *M* classes.

In thresholding technique, the filtering algorithm uses the appropriate range of the threshold parameter Θ which will allow an accurate extraction of the region of interest (ROI) composed of the nuclei and cytoplasm pixels. Threshold technique is detailed as earlier in this survey paper. The parameter Θ is determined by trial and error testing whereby the outcome of the segmentation is assessed visually.

In [6], researchers have provided another technique for segmenting grey level and color images.

Hopfield Neural Network (HNN) is one of the artificial neural networks, which has been proposed for segmenting both gray-level and color images. The HNN is very sensitive to intensity variation and it can detect the overlapping cytoplasm classes. HNN is considered as unsupervised learning. Therefore, the network classifies the feature space without teacher based on the compactness of each cluster calculated using the Euclidean distance measure between the *kth* pixel and the centroid of class *l*.

The HNN segmentation algorithm can be summarized in the following steps [6]:

1) Initialize the input of neurons to random values.
2) Apply the input-output relation given in (8) to obtain the new output value for each neuron, establishing the assignment of pixel to classes.
3) Compute the centroid for each class as follow:

$$\overline{\boldsymbol{xl}} = \frac{\sum_{k=1}^{n} x_k v_{kl}}{n_l} \qquad (2.1)$$

Where, *nl* is the number of pixels in class *l*.

4) Solve the set of differential equation in (7) to update the input of each neuron:

$$\mathbf{U_{kl}(t+1) = U_{kl}(t) + \frac{dU_{kl}}{dt}} \qquad (2.2)$$

5) Repeat from step 2 until convergence then terminate.

Based on the algorithm they concluded that algorithm could segment 97% of the images successfully in nuclei, cytoplasm regions and clear background. Furthermore, HNN took short time to achieve the desired results.

In Fuzzy Clustering approach, The process of clustering is to assign the *q* feature vectors into *K* clusters, for each *kth* cluster *Ck* is its center. Fuzzy Clustering has been used in many fields like pattern recognition and Fuzzy identification. A variety of Fuzzy clustering methods have been proposed and most of them are based upon distance criteria The algorithm has as input a pre-defined number of clusters, which is the *k* from its name. Means stands for an average location of all the members of particular cluster and the output is a partitioning of *k* cluster on a set of objects. The objective of the FCM cluster is to minimize the total weighted mean square error. With the preceding of algorithm they conclude that, FCM is not sensitive to intensity variation, therefore, the cytoplasm regions are detected as one cluster when they fixed the cluster number to three, four, five and six. Moreover, FCM failed in detecting the nuclei, it detected only part of it.

In [6], they concluded that the HNN segmentation results are more accurate and reliable than FCM clustering in all cases. The HNN succeeded in detecting and segmenting the nuclei and cytoplasm regions. However FCM failed in detecting the nuclei, instead it detected only part of it. In addition to that, the FCM is not sensitive to intensity variations as the segmentation error at convergence is larger with FCM compared to that with HNN.

Molecular analysis of sputum has been an active area for the investigation of lung cancer biomarkers for several reasons.

In [7] they discuss various sputum-based molecular biomarkers and how each one has developed during the last decade.

In [7], they defined various sputum based molecular biomarkers. They described cytology, Allelic Alteration, Methylation, Mutation, microRNAs. With the study of this different approach they concluded that standardized sputum collection and processing protocols should be established to minimize the inconsistencies resulting from different laboratories using different methods.

III. METHODOLOGY

For Early detection of Lung cancer using sputum cytology, The main focus of this research is on proper and efficient classification and clustering techniques. With effective method for clustering sputum image, result will be more accurate.

So to extend the derived results of given techniques, it is possible to enhance one of the classification and clustering method and acquire the accurate decision for detection of Lung cancer. There are various classification techniques available for early detection of Lung cancer. In data mining, classification is one of the most important tasks. It maps the data in to predefined targets. It is a supervised learning as targets are predefined. The aim of the classification is to build a classifier based on some cases with some attributes to describe the objects or one attribute to describe the group of the objects.

The most used classification algorithms exploited in the microarray analysis belong to four categories: IF-THEN

Rule, Decision tree, Bayesian classifiers and Neural networks [8].

A. IF conditions THEN conclusion [8]

This sort of lead comprises of two parts. The run the show precursor (the IF part) holds one or more conditions about esteem of indicator qualities whereas the run the show subsequent (THEN part) holds an expectation about the esteem of an objective quality. An accurate prediction of the value of a goal attribute will improve decision-making process.

IF-THEN prediction rules are very popular in data mining; they represent discovered knowledge at a high level of abstraction. Rule Induction Method has the potential to use retrieved cases for predictions [8].

B. Decision Tree: [8]

Decision tree determines from the basic separation and conquer calculation. In these tree structures, leaves speak to classes and limbs speak to conjunctions of characteristics that accelerate those classes. At every hub of the tree, the property that most adequately parts specimens into distinctive classes is picked. To anticipate the class name of an enter; a way to a leaf from the root is found hinging upon the esteem of the predicate at each one hub that is gone to. The most well-known calculations of the choice trees are Id3 and C4.5.

An advancement of choice tree misused for microarray information dissection is the arbitrary timberland, which uses a gathering of arrangement trees.

C. Bayesian classifiers and Naive Bayesian: [8]

From a Bayesian perspective, a characterization issue could be composed as the issue of uncovering the class with greatest likelihood given a set of watched trait values. Such likelihood is seen as the back likelihood of the class given the information, and is generally processed utilizing the Bayes hypothesis. Assessing this likelihood circulation from a preparation dataset is a troublesome issue, on the grounds that it may oblige a quite expansive dataset to fundamentally investigate all the conceivable consolidations.

Alternately, Naive Bayesian is a straightforward probabilistic classifier dependent upon Bayesian hypothesis with the freedom presumption. In light of that administer, utilizing the joint probabilities of example perceptions and classes, the calculation endeavors to gauge the restrictive probabilities of given a perception. Notwithstanding its straightforwardness, the Credulous Bayes classifier is known to be a hearty strategy, which shows on normal exceptional execution regarding characterization exactness, additionally when the autonomy suspicion does not hold.

D. Artificial Neural Networks (ANN) [8]

A simulated neural system is a numerical model taking into account organic neural systems. It comprises of an interconnected aggregation of counterfeit neurons and courses of action data utilizing a connectionist approach to reckoning.

Neurons are arranged into layers. The enter layer comprises essentially of the first ever information, while the yield layer hubs speak to the classes. At that point, there may be some stowed away layers. A key characteristic of neural systems is an iterative taking in process in which information examples are displayed to the system one at once, and the weights are balanced in place to foresee the right class mark. Preferences of neural systems incorporate their high tolerance to boisterous information, too as their capability to group examples on which they have not been prepared.

Based on these basic classification techniques and other more data mining concept, the aim of this research is to find the better approach of these techniques which concludes in detecting lung cancer at early stage.

With Clustering approach, sputum image is converted from RGB to HSV color model. With the help of Morphological operation, sputum image is segmented. That segmented image is clustered into different clusters with the use of appropriate cluster method of data mining. With morphological operation, nuclei of each cluster is detected and then process of detecting abnormal cells is continued based on appropriate approach [4].

IV. CONCLUSION

In this paper, we gone through different classification and clustering techniques of data mining. Different techniques can be used for clustering sputum images based on nucleus detection. Bayesian approach provides the good result for cancer detection but it is not that much proper for sensitivity. It was found that the HNN segmentation results are more accurate and reliable than FCM clustering in all cases. The HNN succeeded in detecting and segmenting the nuclei and cytoplasm regions. However FCM failed in detecting the nuclei, instead it detected only part of it. Hence, there is a need for accurate and more reliable technique which can be useful to detect Lung cancer at early stage with use of Sputum Cytology. Future work, to find the more accurate technique for detecting Lung cancer.

REFERENCES

[1] Kennedy, T.C.; Miller, Y.; Prindiville, "Screening for lung cancer revisited and the role of sputum cytology and fluorescence bronchoscopy in a high-risk group. *Chest J.* **2005**, *10*, 72–79.
[2] Gazdar, A.F.; Minna, J.D. "Molecular detection of early lung cancer". *J. Natl. Cancer Inst.* 1999, *91*, 299–301.
[3] http://www.pamf.org/ teen/health info/
[4] Fatin Izzwani Azman, Kamarul Hawari Ghazali, Rosyati Hamid, Zeehaida Mohamed, Nur Shahida Nawi. "Detection and summation of squamous epithelial cells in sputum slide images by nucleus detection", 4,5Microbiology and Parasitology Department, Health Campus, University Sains Malaysia.
[5] Fatma Taher, Naoufel Werghi, Hussain Al-Ahmad and Christian Donner. "Extraction and Segmentation of Sputum Cells for Lung Cancer Early Diagnosis" ISSN 1999-4893, 2013.
[6] Sammouda , Fatma Taher, Naoufel Werghi, Hussain Al-Ahmad, Rachid "Lung Cancer Detection by Using Artificial Neural Network and Fuzzy Clustering Methods.", American Journal of Biomedical Engineering 2012, 2(3): 136-142
[7] Connie E. Kim, Kam-Meng Tchou-Wong and William N, Rom "Sputum-Based Molecular Biomarkers for the Early Detection of Lung Cancer - Limitations and Promise", *Cancers* 2011, *3*, 2975-2989.
[8] V.Krishnaiah, Dr.G.Narsimha, Dr.N.Subhash Chandra, "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques", IJCSIT, Vol. 4 (1), 2013, 39 – 45.

