Comparative Study Review on Lung Cancer Detection Using Neural Network and Clustering Algorithm

Sukhjinder Kaur

Abstract— **Among various diseases cancer has become major threat in India. As per Indian population due to cancer the mortality rate was very high. Cancer is the second most disease responsible for death in India. In view of these facts, this paper will reviews two techniques i.e. Neural Network (NN) and Fuzzy C Mean Clustering Algorithm with their cons and pros, that are very helpful in early diagnose of lung cancer. In addition to above, in the end this paper will conclude which technique is best and has to be adopted for better accuracy of cancer prevention system.**

***Index Terms*— Lung Cancer, Neural Network (NN), Fuzzy C Mean Clustering Algorithm, sputum images.**

I. INTRODUCTION

The high prevalence of lung cancer leads to its early prevention. The introduction of computer technology helps lot in increasing the mortality rate of the lung cancer patients due to its detection at early stages. Determine whether a pulmonary nodule is a being tumor or not in the early stages is very important. But determination of the presence of tumors in small nodules is very difficult. With the rapid advancement of the technology, the interaction between physics, engineering and computing science has become closer than ever before. More people die because of lung cancer than any other types of cancer such as: Breast, colon, and prostate cancers as shown in Figure.2.Human machine systems for image-based diagnosis need to take advantage of both human and machine capabilities, creating a system, which as a whole will be greater than the sum of its parts *(Katherine et.al, 2003)*.

In India the vast majority cases (90%) of lung cancer is due to exposure to tobacco smoke. About 10 % of cancer occurs in that people who never smoked. These cases are often caused due to the genetic effects. Lund cancer is the most common cause of death in India and was responsible for 1.56 million deaths annually, as per survey in 2012 and in 1991 around 60, 9000 people was effected by lung cancer.

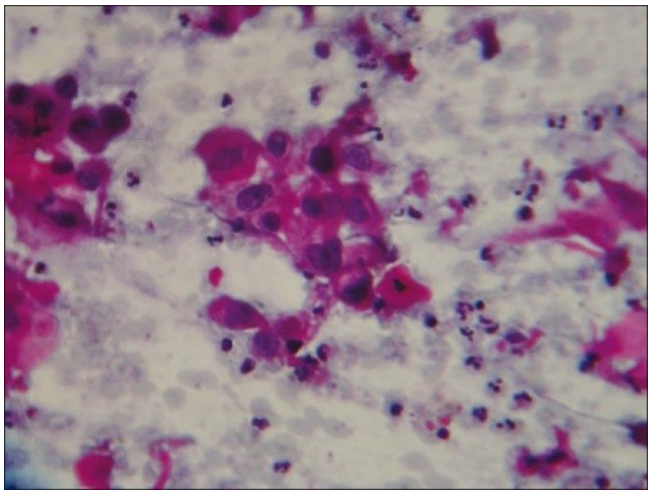

Figure. 1 Sputum color image showing Lung cancer [ ]

Lung cancer staging is an assessment of the degree of spread of the cancer from its original source. It is one of the factors affecting the prognosis and potential treatment of lung cancer *(Hornet.al, 2012)*. Below chart shows the reasons of death in India. From graph it is clearly seen that Lung cancer is at second most place. Recent studies shows that people living at higher altitudes has low risk rate of lung cancer in smoker as well as in non-smoker, which shows that oxygen may promote the lung cancer problem, This is published online in PeerJ. Also according to Oncology Nurse Advisor, with the elevation of the altitudes, lung cancer rate fell to 7.23 cases in 1000 people.

There are many techniques to diagnose lung cancer, such as chest radiograph(x-ra), computed tomography (CT), magnetic resonance imaging (MRI scan) and sputum cytology *(Wang, 2006)*. But most of the techniques are very expensive. Therefore there is great need of a technology in which the mortality rate is very high. A number of medical researchers utilized the analysis of sputum cells for early detection of lung cancer *(Shiela, 2010)*, most recent research relay on quantitative information, such as the size, shape and the ratio of the affected cells *(Kim, 2005)*.

For this reason, most researchers are trying to develop a diagnose system on the basis of sputum color images. There are many algorithms which have been proposed in medical imaging. But in this review paper we will review the most promising methods like Neural Network (NN) and Fuzzy C Mean Clustering Algorithm (FCM) with their advantages as well disadvantages and their methodology to diagnose lung cancer in humans.

ISSN: 2278 – 909X

PER 100,000 POPULATION

GOOD POOR

TOP 50 CAUSES OF DEATH	Rate	World Rank	TOP 50 CAUSES OF DEATH	Rate	World Rank
1. Coronary Heart Disease	165.79	37	26. Pertussis	5.93	19
2. Lung Disease	142.09	1	27. Drownings	5.92	46
3. Diarrhoeal diseases	132.70	11	28. Epilepsy	5.67	43
4. Stroke	116.41	77	29. Oesophagus Cancer	5.34	44
5. Influenza & Pneumonia	68.04	67	30. Maternal Conditions	5.22	56
6. Tuberculosis	28.83	58	31. Fires	5.11	33
7. Hypertension	24.44	107	32. Violence	4.76	100
8. Diabetes Mellitus	23.83	108	33. Congenital Anomalies	4.67	141
9. Liver Disease	23.59	27	34. Endocrine Disorders	4.63	117
10. Falls	23.48	1	35. Hepatitis B	4.15	14
11. Kidney Disease	21.79	66	36. Stomach Cancer	3.98	152
12. Other Injuries	19.93	44	37. Colon-Rectum Cancers	3.18	173
13. Suicide	19.05	16	38. Prostatic Hypertrophy	3.16	1
14. Road Traffic Accidents	18.65	77	39. Alzheimers/Dementia	3.11	116
15. Low Birth Weight	17.15	47	40. Lymphomas	3.02	155
16. HIV/AIDS	16.78	67	41. Syphilis	2.96	22
17. Birth Trauma	13.20	49	42. Leukemia	2.62	145
18. Peptic Ulcer Disease	12.37	5	43. Tetanus	2.50	14
19. Breast Cancer	12.26	147	44. Ovary Cancer	2.27	87
20. Oral Cancer	10.69	8	45. Liver Cancer	2.26	181
21. Cervical Cancer	8.15	62	46. Rheumatic Heart Disease	2.25	58
22. Meningitis	8.07	45	47. Malnutrition	2.09	99
23. Asthma	7.54	71	48. Malaria	1.84	53
24. Measles	7.22	8	49. Hepatitis C	1.68	15
25. Lung Cancers	6.49	136	50. Other Neoplasms	1.61	144

Figure. 2 Causes of Death in India (www.lungindia.com)

The reminder of this paper is organized as follows. In Section 2, Literature survey is presented. In Section 3, Neural Network algorithm is de-scribed. In Section 4, fuzzy clustering algorithm is presented and finally in Section 5, the conclusion and future work are given.

II. RELATED WORK

Fuzzy k-c-means clustering algorithm used for medical image segmentation which was introduced in *(Ajala, 2012).* Here fuzzy-c-means is a method of clustering algorithm which allows one piece of data belongs to two or more clusters and k-means is a simple clustering method in which we use low computational complexity as compared to fuzzy c-means. When both Clustering methods were combined to produce a more time efficient segmentation algorithm called as fuzzy-k-c-means clustering algorithm. They offered that thresholding which is the most elementary technique for medical image segmentation, in which this algorithm divides pixels in different classes depending upon their gray level. It is also said that it approaches division of scalar images by forming a binary partition of the intensity values of an image and lastly determines an intensity value. This intensity value is termed as threshold, which separates the desired classes. Classifier techniques which were used for pattern recognition, partitions a feature space derived from the image using data with known labels. A feature space is a set of N*M matrix where N relates to the number of observations and M relates to the number of attributes. Classifiers are known as supervised methods since they require training data which are manually segmented and then used it for automatically segmenting new data.

A comparison between two methods was made in *(Christian, 2012).* These methods are rule based method and Bayesian classicism method for the extraction of cell region from background and debris cell region, and after experimentation the Bayesian classicism method was found applicin this able for classification of sputum cell region from background region. But they did not remove the nucleus region from cytoplasm region with this technique.

In this *(Fatma, 2012*) two more segmentation methods were used which were Hopfield Neural Network (HNN), and Fuzzy C-Mean (FCM) clustering algorithm. In this they found that the HNN provides enhanced, accurate and reliable segmentation results than FCM clustering in all cases. The HNN also divides the nuclei and cytoplasm regions while FCM failed in the detection of the nuclei. FCM only detected a part of the nucleus not the whole nucleus in a particular cell. Also FCM was not found subtle to intensity variations because the segmentation error at convergence was found larger with FCM in comparison to HNN. According to the utmost latest estimates of the statistics which are provided by world health organization indicates that there happened around 7.6 million deaths worldwide each year because of this type of cancer. Moreover, they also found that mortality from cancer are estimated to rise continuously, and will come near to 17 million deaths worldwide in 2030. So, better methods are required to extract the nucleus region for very early detection. A magazine in *(IEEE, Pulse) provided* us the knowledge about current trends in medical image analysis.

In *(Mokhled, 2012)* first images which were improved through Gabor filter. It has given better results than other enhancement techniques. They only worked on colored image enhancement and not extract the nucleus region and even not the cell region. In Features Extraction stage they acquire the

ISSN: 2278 – 909X

general features of the enhanced and segmented image which later they used in Binarization. A refined Charged Fluid Model (CFM) along with improved Otsu's method was used for the automatic segmentation of MRI images in *(Nagesj, 2012)*. This method gave enhanced results than the result given by the approaches used in previous experiments.

In *(Nikita, 2012),* a sober edge detection method was used which is based on finding the image gradient. This method tells that intensity of the image will be maximum where there is a separation of two dissimilar regions and thus an edge must exist there. On this basis they found the nodules in CT images.

In *(Parsh, 2011),* a new variation level set algorithm without re-initialization was used. They also used thresholding to reduce the noise component of the images.

In *(Sajith, 2012)* glandular cells were detected by using multiple color spaces and two clustering algorithms which were K-means and Fuzzy C-means.

In *(Sonith, 2012)* an overview of entire process for processing digital images for lung cancer detection is given in this paper. This paper also describes all the essential steps required for the better performance starting from the pre-processing till the very end phase extraction of features.

III. NEURAL NETWORK

Machine learning algorithms help a lot in decision making and neural network has performed well in classification purpose in medical field. Most popular techniques among them is neural network. Neural networks are composed of simple elements which operate parallel. A neural network *(V.VThakre et.al, 2010)* can be trained to perform a particular function by adjusting the values of the weights between elements. Network function is determined by the connections between elements. There is activation functions used to produce relevant output.

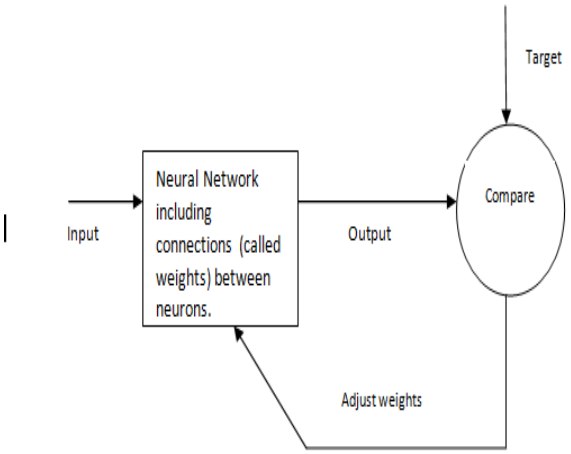

Figure III.1: Neural Net Block Diagram

Training can be either supervised or unsupervised. In supervised training system learns by trying to predict outcomes for known examples. System compares its predicitions with the known results and learns from is mistakes. In unsupervised training system no output or result is shown as part of training process. With the delta rule, as with other types of back propagation, 'learning' is a supervised process that occurs with each cycle or 'epoch' (i.e. each time the network is presented with a new input pattern) through a forward activation flow of outputs, and the backwards error propagation of weight adjustments. Simply, when a neural network is initially presented with a pattern it makes a random 'guess' as to what it might be. It then sees how far its answer was from the actual one and makes an appropriate adjustment to its connection weights. Within each hidden layer node is a sigmoidal activation function which polarizes network activity and helps it to be stable in nature. Back propagation performs a gradient descent within the solution's vector space towards a 'global minimum' along the steepest vector of the error surface. The global minimum is that theoretical solution with the lowest possible error. Back propagation is a method of training artificial neural network. It requires a desired output for each value in order for calculation of loss function gradient. Following algorithm will show how BPNN works in classification in medical imaging.

Basically the error back-propagation process consists of two passes through the different layers of the network a forward pass and a backward pass.

The algorithm is as follows:
(1) Step 0. Initialize weights. (Set to small random values)
(2) Step 1. While stopping condition is false do steps 2–9
(3) Step 2. For each training pair do steps 3–8
Feed-forward
Steps 3: Each input unit (Xi, i = 1 . . . n) receives input signal Xi and sends signal to all units in the layer above (the unseeable units).
Steps 4: Each hidden unit (Zj, j = 1, p) sums its weighted input Signals,

$\mathbf{Z_{in}= V_{oj}+\sum^{n}_{i=1} X_i V_{ij}}$ V0j : Bias on hidden unit j.

Vij : Weight between input unit and hidden unit. Applies its activation function to calculate its output signal.

$\mathbf{Z_j=f(Zin_j)}$

And conveys this signal to all units in the layer above (output units).

Steps 5:Each output unit (Yk, k = 1, . . . , m) sums its burthened input signals.

$\mathbf{Yin_k = W_{ok}+\sum^{p}_{j=1} Z_j W_{jk}}$

Wok: Bias on output unit k.
Wjk :Weight between hidden unit and output unit.
And uses its activation function to compute its output signal,

$\mathbf{Y_k = f(yin_k)}$

B. Back propagation of error

Step 6: Each output unit (Yk,k = 1, . . . , m) receives a target pattern matching to the input training pattern, calculates its error information term,

ISSN: 2278 – 909X

$\delta_k = (t_k - y_k) f(yin_k)$
calculates its weight correction term (used to update Wjk later),
$\Delta Wjk = \partial \delta_k Z_j$
Calculates its bias correction term (used to update Wok later),
$\Delta W_{ok} = \partial \delta_k$
And sends δk to units in the layer beneath.
Step 7.: Each hidden unit (zj, j = 1, . . . , p)adds up its delta inputs (from unit in the layer above),
$\partial in_j = \sum_{k=1}^{m} \partial_k Wjk$
Multiplies by the derivative of its activation function to calculate its error information term,
$\partial_j = \partial inj f(Zinj)$
Calculates its weight correction term (used to update Vij later),
$\Delta Vij = \partial \delta j Xi$
And calculates its bias correction term (used to update Voj later),
$\Delta Voj = \partial \delta j$
Update weights and biases:
Step 8: Each output unit (Yk, k = 1, . . . , m) updates its bias and weights (j =. . . p) :
Wjk(new)=Wjk(old)+ΔWjk
unit (Zj, j = 1, . . . , p) updates its bias and weights (i = 0, . . . , n):
Vij(new)=vij (old)+ΔVij
Step 9: Test stopping condition.

Advantages:

- They are good at adapting to changing situations.
- Neural network build models that are more reflective of the structure of the data in significantly less time.
- Neural networks operate well with modest computer hardware.
- A neural network can continue without any problem even if an element of neural network fails.

Disadvantage:

- Key limitation of neural network is its inability to explain how the network has been build. Neural network gets better answer but have hard time explaining how they got there.
- Extraction of rules from neural network is difficult.
- Time consuming process of training the neural network from complex data set.

Neural network needs training to operate.

IV. CLUSTERING ALGORITHM

Clustering is the process of separating the data into identical regions based on the resemblance of objects; information that is logically related physically is stored together, in order to rise the efficiency in the database system and to minimize the number of disk access *(R.Duda, 2001)*. The process of clustering is used to assign the *q* feature vectors into *K* clusters, for each k^{th}cluster $C^k$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 suggested and most of them are based upon distance criteria (*Ramesh, 2011)*. The most extensively used algorithm is the Fuzzy C-Mean algorithm (FCM), because it uses reciprocal distance to compute fuzzy weights. This algorithm has an input a pre-defined number of clusters, which is the *k* from its name. Here 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 main objective of the FCM cluster is to minimize the total weighted mean square error (*Sun, 2004):*
Formula $J = (W^{qk}, C^{(k)}) = \sum \sum (W^{q}_{k})^{p} \left\| x^{(q)} - c^{(k)} \right\|^2$

The FCM allows individually every feature vector to belong to multiple clusters using various fuzzy membership values. Then the final classification will be according to the maximum weight of the feature vector over all clusters. The detailed algorithm (Sun, 2004):
Input: Vectors of objects, each object represent *s* dimensions, where $v=\{v_1, v_2, \ldots\ldots, v_n\}$ in our case it will be an image pixels, each pixel has three dimensions RGB, *K* = number of clusters.
Output = A set of *K* clusters which minimize the sum of distance error.
1. Initialize random weight for each pixel, it uses fuzz weighting with positive weights $\{W^{qk}\}$ between[0, 1].
2. Standardize the initial weights for each q^{th} feature vector over all K clusters via:
W_{qk} / W_{qr}
3. Standardize the weights over k = 1,…,K for each q to obtain W_{qk}, via: (*R.Duda, 2001)*
$W_{qk} = W_{qk} / \sum W_{qr}$, q = 1,…Q.
4. Compute new centroids $C^{(k)}$, k = 1,….,K via
$C^{(k)} = \sum W_{qk} X^{(q)}$, k=1,….K
5. Update the weights { W_{qk}, }
6. If there is any change in the input, repeat from step 3, or else terminate.
7. Assign each pixel to a cluster based on the maximum weight.

Advantage:

- Gives best result for overlapped data set.
- Data point is assigned membership to each cluster center as a result of which data point may belong to more than one cluster center.

Disadvantage:

- Apriori specification of the number of clusters.

ISSN: 2278 – 909X

- With lower value of β we get the better result but at the expense of more number of iteration.
- Euclidean distance measures can unequally weight underlying factors.

V. CONCLUSION AND FUTURE SCOPE

The early detection of lung cancer is a challenging problem, due to the structure of the cancer cells, where most of the cells are overlapped with each other. This paper has presented two segmentation methods, Neural Network (NN) and a Fuzzy C-Mean (FCM) clustering algorithm, for sputum color images to detect the lung cancer in its early stages. The manual analysis of the sputum samples is time consuming, inaccurate and requires intensive trained person to avoid diagnostic errors. This paper also concludes that which method is best along with its advantages so that further work can be done using Neural Network in preference to Fuzzy C-Mean (FCM) clustering algorithm. As Fuzzy C-Mean (FCM) clustering algorithm is not good at low intensity variations.

REFERENCES

[1] Katherine P. Andriole, "Addressing thc Corning Radiology Crisis: The Society for Computer Applications in Radiology, Transforming the Radiological Intcrprctation Process (TRIP.) initsativc."Position Paper from the SCAR TRIPTM Subcommittee of lhc SCAR Research and Development Committee, November 2003.

[2] www.lungindia.com.

[3] Horn, L; Pao W; Johnson DH (2012). "Chapter 89". In Longo, DL; Kasper, DL; Jameson, JL; Fauci, AS; Hauser, SL; Loscalzo, J. Harrison's Principles of Internal Medicine (18th ed.). McGraw-Hill.

[4] http://www.worldlifeexpectancy.com/country-health-profile/india.

[5] W. Wang and S. Wu, "A Study on Lung Cancer Detection by Image Processing", proceeding of the IEEE conference on Communications, Circuits and Systems, pp. 371-374, 2006.

[6] A. Sheila and T. Ried "Interphase Cytogenetics of Sputum Cells for the Early Detection of Lung Carcinogenesis", Journal of Cancer Prevention Research, vol. 3, no. 4, pp. 416-419, March, 2010.

[7] D. Kim, C. Chung and K. Barnard, "Relevance Feedback using Adaptive Clustering for Image Similarity Retrieval," Journal of Systems and Software, vol. 78, pp. 9-23, Oct. 2005.

[8] Ajala Funmilola A, Oke O.A, Adedeji T.O, Alade O.M, Oyo Adewusi E.A, "Fuzzy k-c-means Clustering Algorithm for Medical Image Segmentation", Journal of Information Engineering and Applications, ISSN 2224-5782 (print) ISSN 2225-0506 (online), Vol 2, No.6, 2012 .

[9] Christian D., Naoufel W., Fatma T., Hussain, "Cell Extraction from Sputum Images for Early lung Cancer Detection", IEEE 978-1-4673-0784-0/12, 2012 .

[10] Fatma T., Naoufel W., Hussain, Rachid S., "Lung Cancer Detection by Using Artificial Neural Network and Fuzzy Clustering Methods", American Journal of Biomedical Engineering, 136-142 DOI: 0.5923/j.ajbe.20120203.08, 2012 .

[11] "Medical Image Analysis", IEEE Pulse, 2154-2287/11/2011.

[12] Mokhled S. AL-TARAWNEH, "Lung Cancer Detection Using Image Processing Techniques", Leonardo Electronic Journal of Practices and Technologies, ISSN 1583-1078, Issue 20, January-June 2012 .

[13] Nagesh V., Srinivas Y., Suvarna Kumar G, Vamsee Krishna V, "An Improved Medical Image Segmentation Using Charged fluid Model", International Journal of Engineering and Applications (IJERA) ISSN: 2248-9622, Vol. 2, Issue 2, pp.666-668, Mar-Apr 2012 .

[14] Nikita P., Sayani N., "A Novel Approach of Cancerous Cells Detection from Lungs CT Scan Images", International Journal of Advanced Research in Computer Science and Software Engineering, ISSN 2277 128X, Volume 2, Issue 8, August2012 .

[15] Parsh Chandra B., Md. Sipon M., Bikash Chandra S. and Mst. Tiasa K., "MRI Image Segmentation Using Level Set Method and Implement a Medical Diagnosis System", Computer Science & Engineering: An International Journal (CSEIJ), Vol. 1, No. 5, December 2011 .

[16] Sajith Kecheril S, D Venkataraman, J Suganthi and K Sujathan, "Segmentation of Lung Glandular Cells using Multiple Color Spaces", International Journal of Computer Science, Engineering and Applications (IJCSEA) Vol.2, No.3, June 2012 .

[17] Sonit Sukhraj Singh, Anita Chaudhary "Lung Cancer Detection using Digital Image Processing", IJREAS Volume 2, Issue 2 ISSN: 2249-3905, (February 2012) .

[18] V.V. Thakare, P. Singhal," Neural network based CAD model for the design of rectangular patch antennas, " JETR,vol. 2(7), 2010.

[19] R. Duda, P. Hart, "Pattern Classification", Wiley-Interscience 2nd edition, October 2001.

[20] S. Aravind, J. Ramesh, P. Vanathi and K. Gunavathi," Rou-bust and Automated lung Nodule Diagnosis from CT Images based on fuzzy Systems", processing in International Conference on Process Automation, Control and Computing (PACC), pp. 1-6, Coimbatore, India, July, 2011.

[21] H. Sun, S. Wang and Q. Jiang, "Fuzzy C-Mean based Model Selection Algorithms for Determining the Number of Clusters," Pattern Recognition, vol. 37, pp.2027-2037, 2004S.

ISSN: 2278 – 909X

Sukhjinder Kaur received the B.Tech degree in Computer Science Engineering from Guru Gobind Singh College of Engineering and Technology, Kharar (Punjab) in 2011. Now she is pursuing M.Tech in Computer Science Engineering from Shaheed Udham Singh Group of Institutions, Tangori (Punjab) and Working as a Lecturer in Global College of Engineering and Technology, Anandpur Sahib (Punjab). Her research area is Digital Image Processing.

ISSN: 2278 – 909X