

Entropy Based Lung Cancer Prediction

¹Dimpy Raghav, ²Priyanka Srivastava, ³Nancy Singh ⁴Harsh Rawat

Department of Information Technology, IPEC, Sahibabad,U.P. , India¹⁻⁴

Abstract: We all know about various types of hazardous diseases but out of them all the Cancer is the most fatal and common among people. A large number of population get suffered and lose their lives due to it. If cancer is diagnosed in early stages it could be cured, but if it diagnosed in later stages, the chances of survival became negligible. The prominent cause of cancer-related mortality throughout the world is "Lung Cancer". Hence beforehand detection, prediction and diagnosis of lung cancer is a necessity as it can increase the chances of survival. Various types of machine learning algorithms (ML) like Naive Bayes, Support Vector Machine (SVM), Logistic regression, Artificial Neural Network (ANN), Convolutional Neural Network (CNN) have been applied in the healthcare sector for analysis and prognosis of lung cancer. This paper will highlight the methods by which we can diagnosis or predict the presence of the tumor in the lungs using image data.

Keywords: Lung Cancer, Entropy, Classification, Thresholding, Tumor, Histogram, Segmentation, Dilation.

I. INTRODUCTION

Cellular breakdown in the lungs is an unsafe sickness that causes an enormous number of passings worldwide. The base experience of cellular breakdown in the lungs is important to diminish the death pace of patients. Thus it is an incredible test experienced by specialists and scientists to distinguish and analyze tumors in the lungs. Throughout the last many years, an unremitting improvement that relates to the disease research has been proposed to a serious degree. Different exploration works have executed various models for the previous acknowledgment of disease prior to experiencing signs. By the creation of new models in clinical regions, enormous disease information are assembled and are openly available by the clinical examination society. Be that as it may, there is one critical moving errand to doctors for example the infection ought to be anticipated precisely. Recognition of cellular breakdown in the lungs should be possible by utilizing clinical pictures like registered tomography, chest X-beam; MRI checks, and so on, ML approaches perceive the primary qualities of mind boggling cellular breakdown in the lungs datasets. A CAD (Computer-Aided Diagnosis) was created in the mid 1980s to upgrade the endurance rate and effectiveness that help the specialists in deciphering clinical pictures. Some of the AI calculations that have a significant effect in medical care are choice trees, direct relapse, irregular woodland, SVM, gullible Bayes, K-closest neighbors, etc. We have likewise examined the profound learning strategies methods and calculations that can be executed for conclusion, location, and prediction of various cancers. The preeminent intent of this research work is to introduce a succinct vision of present work on cellular breakdown in the lungs expectation utilizing profound learning and AI models.

Symptoms are categorized based on the location and size of the tumor [35]. During the early stages, it's difficult to analyze and detect as it will not any cause any pain and symptoms in some cases. Lung cancer diagnosed patient may suffer through Cough, Chest pain, Shortness of breath, Wheezing, Hemoptysis i.e. coughing up blood, Pancoast syndrome (shoulder pain), Hoarseness (paralysis of vocal cords), Weight loss, Weakness, and Fatigue. Types of lung cancer are pictorially shown in figure 1.

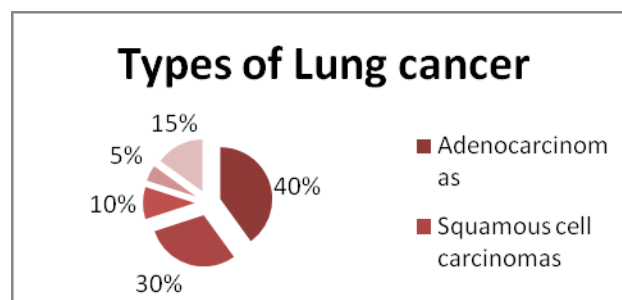


Figure 1: Types of lung cancer

90% of it is induced due to smoking. Impregnation of tobacco smoke also causes lung cancer i.e. known as passive smoking. Another factor for lung cancer is heredity. Vehicular pollution, industries, the intake of harmful gases such as



Radon stands at the second position in causing the deaths from lung cancer..Factors causing lung cancer and mortality rate are shown in table 1.

Table I. Factors Causing Lung Cancer and Mortality Rate

Causes	Mortality in%	Figure
Cigarette smoking□	90%	70000 (USA)
Radon gases	12%	21000
Passive smoking□	2-4%	-----

In this research paper we will create a method by which using machine learning we can diagnose easily if a person is affected or not by using some images datasets. Using this method one person can easily identify the situation of the patient.

Table II. LUNG CANCER RATE ALL OVER THE GLOBE TABLE III LUNG CANCER CASES AND DEATHS IN INDIA

Region	Population	Cancer Cases	Deaths
Asia	60%	66%	57.3%
Europe	9.0%	23.4%	20.3%
America	13.3%	21.0%	14.4%
Africa	17%	9-10%	7.3%

Lung Cancer	New Cases	Deaths
Men	60%	66%
Women	9.0%	23.4%
Both cases	13.3%	21.0%

II. LITERATURE REVIEW

Various sorts of AI procedures have been created to further develop the analysis exactness pace of cellular breakdown in the lungs. For that, as needs be CAD frameworks are introduced. S. Ashwin, et al. [6] and Ada, et al. [7] introduced CAD framework in light of Artificial Neural Network for the lung knob identification. Their proposed frameworks are given precision 96.7% and 96.03% separately. In 2015, K.Punithavathy et al. [4] made sense of cellular breakdown in the lungs recognition in view of surface highlights and Fuzzy C means. The paper mostly focuses on the picture pre-handling parts utilizing various strategies to obtain improved results and a grouping technique to produce the result. In the pre-handling part, to build the differentiation present in the Computed Tomography Images (CT pictures), Contrast Limited Adaptive Histogram Equalization (CLAHE) was applied. Rather than applying this method to the entire picture, it is applied to little districts of the pictures known as tiles. Bilinear addition is utilized to consolidate the different improved parts/districts of the picture. Wiener channels are utilized to diminish the commotion by a critical sum. Locale extraction assumes a significant part to get the ideal area. Morphological activities, for example, shutting were utilized to get the ideal area for example area having lung projections and abandoning the veins, bronchi, and any remaining inward parts. The organizing component of plate shape was utilized in the end activity. While in the component extraction process, surface based highlights were concentrated as force esteem isn't the right boundary to extricate highlights.

The characterization of the pre-handled picture is finished utilizing FCM. Anam Tariq, et al. [8] introduced a proposed strategy in view of Neuro-Fuzzy classifier for finding of cellular breakdown in the lungs which gives framework exactness 95%. Utilizing nearby double strategy, for malignant growth finding, Yeni Hardi, et al. [9] introduced probabilistic brain network which has precision 78% for 3D and 43% for 2D. Fatma Taher [10]-[11] introduced two proposed strategy, one depends on Bayesian order and other one depends on Artificial Neural Network and Fuzzy Clustering technique for finding of cellular breakdown in the lungs having precision of framework is 88.6%.

Hamada R.H., et al. [12] proposed CAD structure considering K-nearest neighbor classifier which has precision 96.5%.. The maker [13] has completed CAD structure using SVM, which gives the accuracy 96.6 % on informational collection of 150 models for recognizable proof of cell breakdown in the lungs. The essential objective of our proposed work is to remove ideal particular components from lung handle, which will help the CAD system to review the accuracy of classifier ANN with SVM for cell breakdown in the lungs assurance. Since the ongoing structure [13] has managed 150 models. In the proposed system, 250 models are used and in this way precision rate using SVM and ANN has been examined. The proposed structure is made from following different stages; Database, Pre-taking care of, Feature extraction, Feature decision and Classification. Data Base The proposed CAD structure has worked on complete 250 lung CT pictures, out of which 125 models are disastrous cases with single and various handles and other 125 models are customary lung pictures. Right around 190 models are accumulated from site cancerimagingarchives.net and the rest of tests are from Mahatma Gandhi Mission and Tapadia Diagnostic Center organized at Aurangabad. All photos are in



JPEG structure with 512*512 pixels objective. The size of handles in informational collection is going from 3.7 mm to 17.2 mm in estimation. The models for picking the size of lung handle/mass relies upon [14]. Pre-Processing The objective of picture pre-dealing with is flawlessly, overhaul and division. Due to pre-dealing with strategy, the idea of CT lung picture will be improved and underline explicit features that makes features extraction/assurance and request all the more remarkable. In proposed structure, pre-taking care of incorporates following advances. De-noising Improvement in the idea of ruined pictures in light of the effect of CT result can be achieved by using utilization of different redesign strategies. There are different sorts of disturbances that degenerate the image ex. Added substance disturbance, Poisson fuss, Gaussian upheaval, etc. To perceive the developments or sickness, the edges ought to be saved. In proposed system, we have involved un-sharp covering to highlight fine nuances inside an image.

In 2019, Moradi et al. [27] contrasted various methods with separate cellular breakdown in the lungs knobs from non-knobs. To lessen/dispose of the misleading positive forecasts they have concocted 3D Convolutional Neural Network Technique. Knobs exist in various sizes and utilizing only one CNN can bring about bogus discoveries. So they separated the knobs into four gatherings as indicated by their size. Also, they have utilized four unique sizes of 3D CNN. They joined that large number of 4 classifiers to come by improved results. Each CNN comprises of various 3D CNN which are changing sizes. Every one of the 4 classifiers were consolidated to create results which were better .

2020, S. Shanthi et al. [22] proposed a framework comprising of a stochastic dispersion search calculation (SDS) and grouping calculations, for example, Neural organizations, Decision trees, and Naive Bayes to recognize cellular breakdown in the lungs. 270 pictures (140 typical and 130 strange) from a dataset named TCGA were gained and utilized. Dark level co-event lattice (GLCM) was applied in order to remove the elements of surface. The Gabor channel was utilized for shape-based highlights. SDS calculation was utilized for highlight determination. It has fundamentally 4 stages - Initialisation stage (task of specialists to a few irregular speculations), Evaluation phase (evaluate the fitness value to find the maximum), Test Phase (Active Agent if: current agent's fitness value > random agent's fitness value, in any other case Inactive Agent) and Diffusion phase (select a random agent if the current agent is inactive else copy the hypothesis of the current agent and offset it). After applying SDS, different classification methods were applied. After observing the accuracies of all the classification models, Neural Network along with the SDS algorithm (SDS-NN) proved to perform better as compared to others. An observation was made implying that the classification of images improves with improved feature selection

ZhiPeng Guo, Yi Xin & YiZhang Zhao (2018) Cancer classification using entropy analysis in fractional Fourier domain of gene expression profile,. This study combined the fractional Fourier transform (FRFT) and entropy-based technique to analyse the gene expression profile (GEP). First, the raw data were transformed into a special fractional Fourier domain with a selected order of FRFT, which had been verified for suiting the pattern recognitions of biological signals as well as for reducing noise [8]. Next, FRFT was combined with an entropy-based method, which could extract inherent genetic features on a genome-wide scale. Finally, tumor classification was performed using the support vector machine (SVM). This method offers a number of advantages; for example, this algorithm has the ability to classify cancers into various subtypes with high accuracy and reflect the inherent relativity of gene ne The entropy method is a kind of objective weighting method; it calculates the weight of indexes by entropy. Subjective fixed-weight methods such as the Delphi method are usually used when determining weights of indexes [19,20]. Such methods could lead to subjective deviations. The entropy weight method is an objective fixed-weight method, which utilizes the quantity of information to determine the weight of an index of interest. Such methods are based on the nature of indexes to determine their weights, which could eliminate subjective deviations and ensure that the results are more concurrent with the facts.

METHODOLOGY

There are a number of techniques that can be used to detect and classify the tumor cells in the image , but first we need to do the steps before operating the image for prediction and detection. In recent times, to use computer technology to solve this problem, several computer-aided diagnosis (CAD) techniques as well as system have been proposed, developed as well as emerged. Those systems use various Machine learning techniques as well as deep learning techniques, there also have been several methods based off of image processing-based techniques to predict the malignancy level of cancer

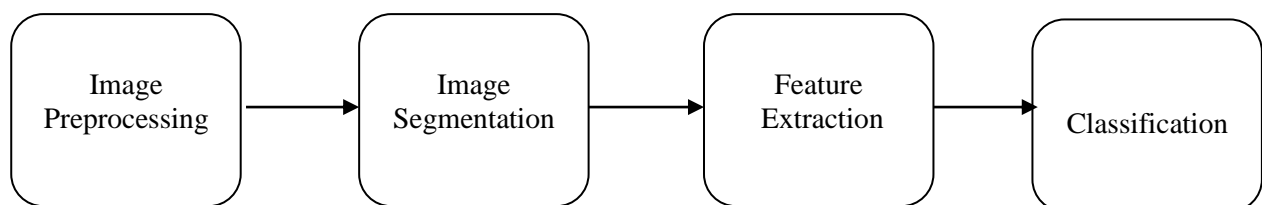
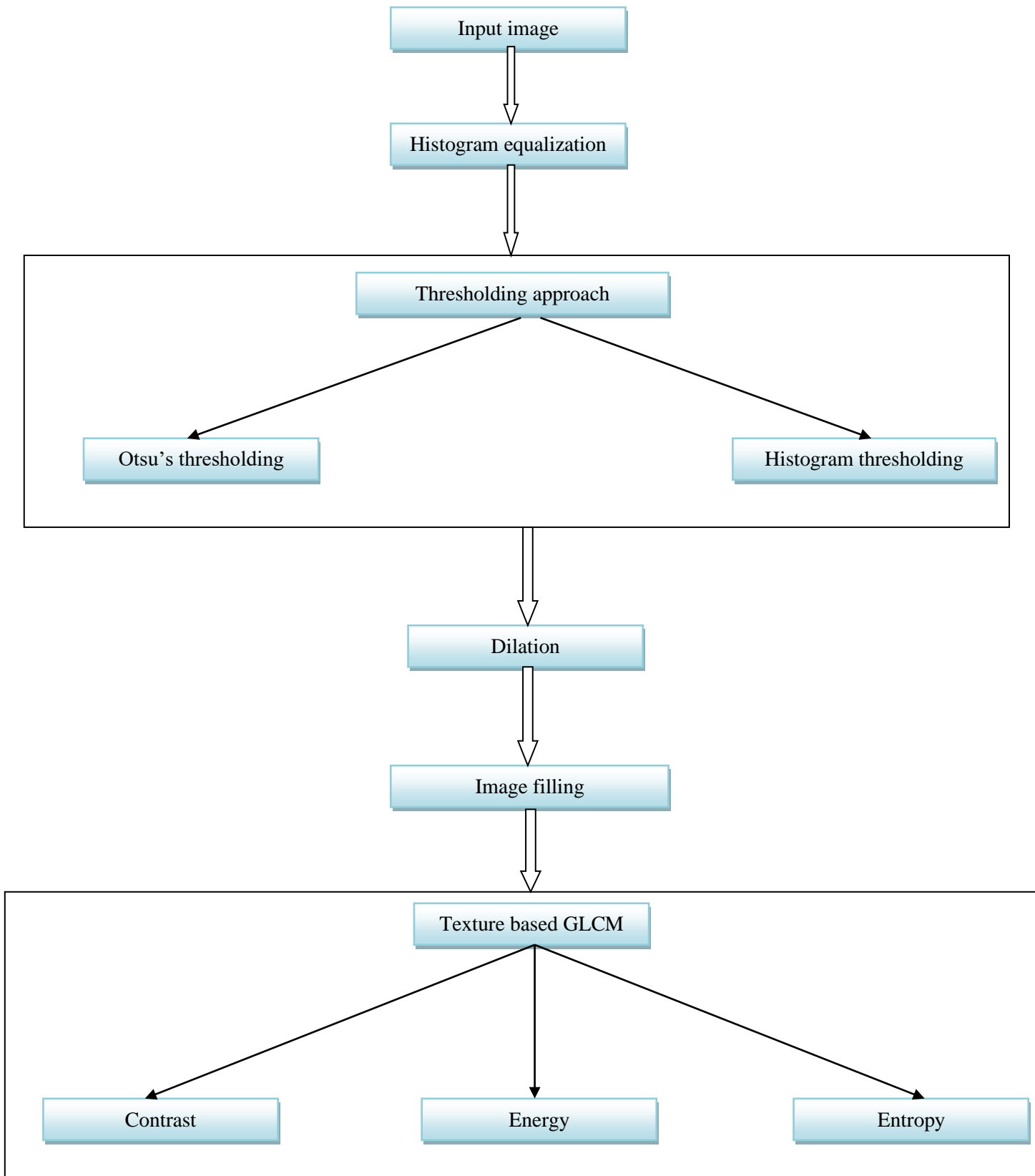


Image Pre-Processing



PROCESS FLOW:



A CAD framework can't straightforwardly utilize CT pictures. They should be well pre-handled before the real use.



Different Image pre-handling procedures are utilized to dispose of commotion and to make pictures appropriate for use. This aides in the improvement of the presentation of the entire framework and consequently the precision.

Image Segmentation

The strategy for dividing a picture into a few portions is known as picture division. Division of picture is done significantly to track down limits in the given picture. The method involved with examining the picture becomes simpler as division decreases the picture intricacy

Feature Extraction

Highlight Extraction is a strategy by which we target decreasing the quantity of aspects that our crude information contains so it is simpler to process and is in a type of sensible classes. Factors in an immense number requiring computational assets to process and create results are trademark for the enormous measures of information. Highlight Extraction procedures manage working on the information while simultaneously guaranteeing that no information is lost. These methods are answerable for picking and consolidating the highlights to limit how much information.

Image Classification

Grouping of pictures is an essential errand that tries to decipher an image overall. By doling out it to a specific mark, the intention is to distinguish the picture. Picture Classification for the most part alludes to pictures where just a single item shows up and is inspected. Object recognizable proof, then again, requires both characterization and restriction assignments and is utilized to look at additional pragmatic cases in which a picture might have a few items. These are the four basic steps in the process of detection of a disease by its diagnosis image . The process flow of the project includes preprocessing of image and then analyzing the results of the medical image to get our results .

This above mentioned flowchart is the representation of the proposed work. The steps and the processes to implement these steps are mentioned below:

Input Image: In this step we select a image and apply it for classification through which we get all pre-processing images.

Image Enhancement: The second we have applied is image enhancement in this step filters are being applied filters are on the images to remove some problems of images such as noise, blurring and etc. For image enhancement different types of filters are being applied on images, here we are applying filters like Historical equalization.

Histogram Equalization: Histogram equalization is the one of the well-known methods for enhancing the contrast of given images in accordance with the sample distribution of an image.

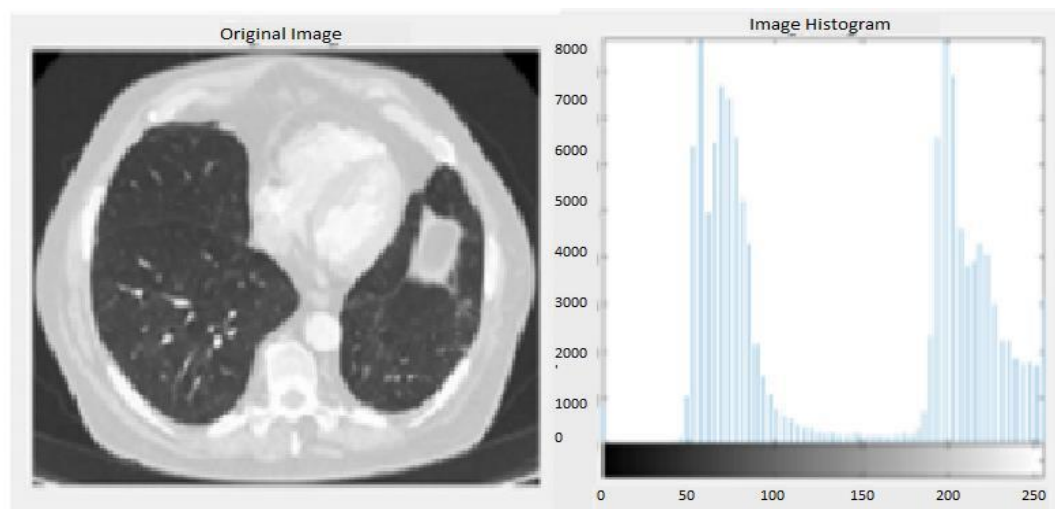


Fig: 2

Segmentation: Segmentation is a important step in image processing. Through the help of segmentation images are divided to some regions that contents of each regions have the same specification.

Thresholding: In this project we are applying thresholding approach which is one of the most powerful tools for image



segmentation. The image obtained from thresholding has the advantages of smaller storage space, fast processing speed of easy in manipulation.

Dilation Operation: Dilation operation is used to extract image component which is used to extract image components that are useful in the representation and description of region shape such as boundaries, skeletons and convex hull.

Feature Extraction: In feature extraction there are several methods through which we can detect or remove portions that are present in a image. To analyze the probability of lung cancer presence, we are applying- Gray level co-occurrence matrix.

GLCM: GLCM is Grey Level Co-occurrence matrix. It is a matrix where the no. of rows and columns is equal to the number of gray level in the image.

Entropy: Entropy is the statistical measure of randomness that can be used to characterize the texture of the input image.

RESULTS

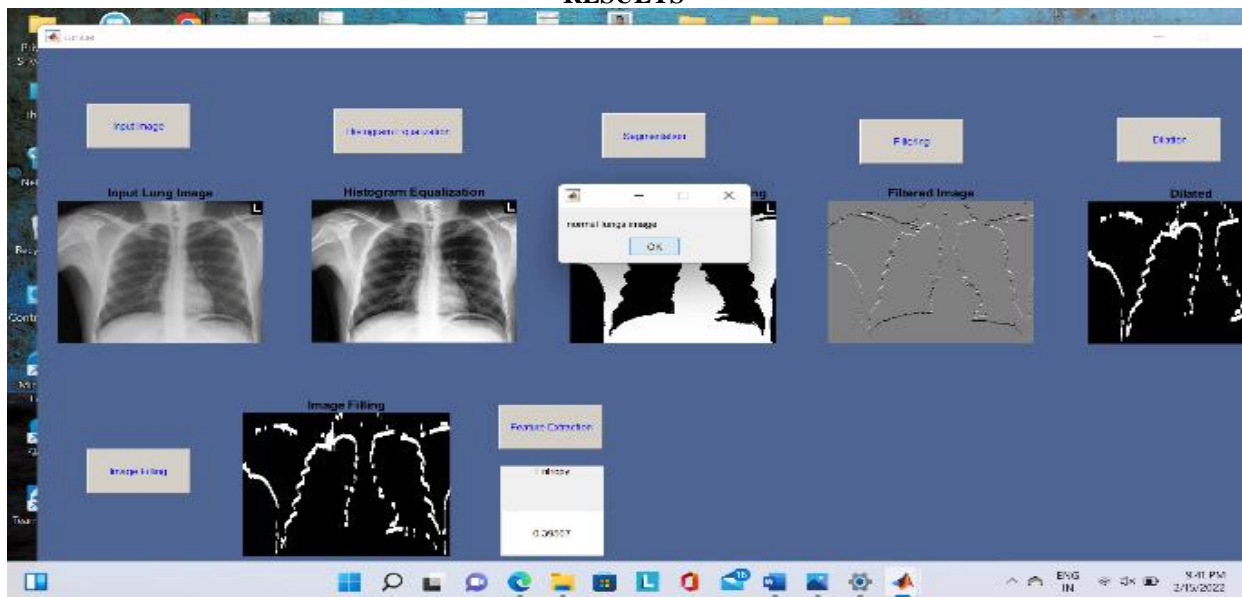


Fig: 3

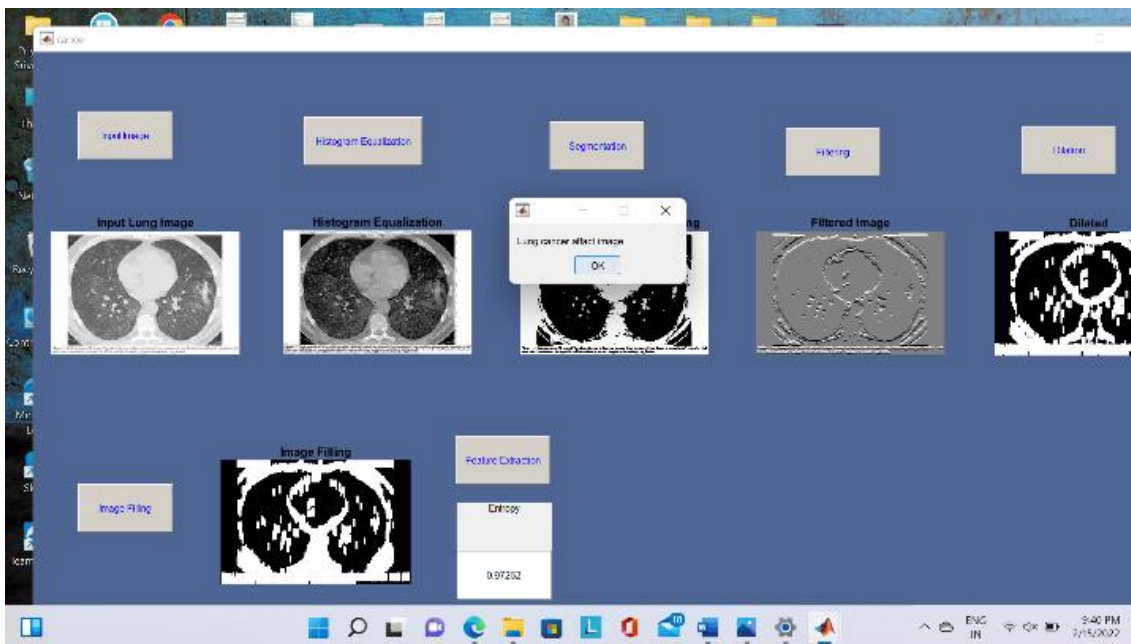


Fig:3.2

It can be clearly observed that in the first figure the calculated results are representing the normal lungs image and in



the second image the image is of lung cancer affected person .The images are preprocessed in the initial steps and then it is analyzed and then the entropy is calculated which will help us to predict the presence of tumor cells. This project will give the results easily and more accurately than the previously known approaches.

COMPARISON WITH OTHER METHODS

In this project we have used Feature extraction method based on entropy measure. Feature extraction based on Entropy is described in two innovative and different ways. Pattern recognition for different types of signals, in 1D and 2D, benefit from the proposed approaches. The proposed methods, which are based on entropy, are independent of the classifiers. There are number of methods which can be used for the detection of the cancerous image like Convolutional Neural Network, Support Vector Machine ,Artificial Neural Network Computer Aided Diagnosis and so on , out of all them this entropy based feature extraction methods is better in a number of ways . The first reason is because it very convenient to use as it is user friendly and less complex as compared to other existing methods .It is cost effective .Any medical professional even with slightest of computer knowledge can operate it . Although this system is concerned with the stage detection but it gives accuracy upto 80-90%. Its user interface is very cooperative to everyone to use as it doesn't concern the unwanted and complex information about the implementation, it just predicts the presence the tumor cells with no unwanted knowledge. Due to all these reasons, this method is a better and easy way to predict the presence of the cancer cells in the lungs.

CONCLUSION

The presented work is the detection of lung cancer nodules by applying implementation on image pre processing and segmentation. By implementing these steps the nodules are detected and then some features are extracted. Then the obtain features are used for the detection. After that we have applied prediction model by applying that we predict from the obtained dataset from feature extraction to know how many people suffering from cancer or not. This technique helps the radiologists and the doctors by providing more information and taking correct decision for lung cancer patient in short time with accuracy. Therefore, this method is less costly, less time consuming and easy to implement. This model is entropy based and focus on finding the value of entropy and thus after processing the image i.e. either x-ray or CT-scan this system tells us the calculated value of entropy .With the help of this system anyone can easily detect the presence of the cancer.

REFERENCES

- [1] World health organization, <http://www.who.int/mediacentre/factsheets/fs297/en/>
- [2] American Cancer Society <http://www.cancer.org/research/cancerfactsstatistics/cancerfactsfigures2014/>
- [3] American Lung Association <http://www.lung.org/lungdisease/lung-cancer/resources/facts-figures/lung-cancerfact-sheet.html>
- [4] American Cancer society <http://www.cancer.org/cancer/lungcancer/>
- [5] Mayo clinic, "Lung Cancer" <http://www.mayoclinic.com/health/lungnodules/AN01082>.
- [6] S. Ashwin, J. Ramesh , "Efficient and reliable lung nodule detection using NN based CAD system". IEEE, ICETEEEM, PP 135- 142, 2012
- [7] Ada, RajneetKaur , " Early Detection and Prediction of Lung Cancer Survival using Neural Network Classifier", IJAIEEM, Volume 2, Issue 6,PP 375-383,June 2013
- [8] AnamTariq,M. Usman , " Lung Nodule Detection in CT images using neuro fuzzy classifier". IEEE , CIMI, PP 49-53, 2013.
- [9] Yeni Hardiyeni,"Diagnosis of lung cancer using 2D and 3D local binary pattern". IJACSA, Vol 3, No. 4, PP 89-95, 2012.
- [10] Fatma Taher , "Bayesian classification and ANN for diagnosis of lung cancer", IEEE, PP 773-776, 2012 .
- [11] Fatma Taher , "Lung Cancer Detection by Using Artificial Neural Network and Fuzzy ClusteringMethods", American Journal of Biomedical Engineering, PP 136-142, 2012.
- [12] Hamada R. H., A- Absi , " CAD System Based on M/C Learning Techniques for Lung Cancer", IEEE , ICCIS, PP 295-300, 2012.
- [13] RashmeeKohad, Vijaya Ahire, "Diagnosis of Lung Cancer Using Support Vector Machine with Ant Colony Optimization Technique", IJACST, Vol.3, No.11, Pages:19-25 (2014)
- [14] Radiology assistant, <http://www.radiologyassistant.nl/en/p460f9fcd50637/solitarypulmonary-nodule-benign-versus-malignant.html>
- [15] Anjali Gautam, H.S. Bhadauria," White Blood Nucleus Segmentation Using an Automated Thresholding and Mathematical Morphing",ICAET-2014.
- [16] Robert M. Haralick," Texture Features for Image Classification", IEEE Transaction on systems, MAN And Cybernetics, PP 610-621,November 1973.



- [17] Ling Chen, Bolun Chen, Yixin Chen, Image Feature Selection Based on Ant Colony Optimization .
- [18] The United States of America. Library of Congress Cataloging-in-Publication Data . Dorigo, Marco. Ant colony optimization / Marco Dorigo, Thomas Stützle. p. cm.
- [19] P.Thukaram,” Image Edge Detection Using Improved Ant Colony Optimization Algorithm”, International Journal of Research in Computer and Communication Technology, PP 1256-1260, Vol 2,Issue 11, November2013
- [20] Bottou, L., and Chih-Jen Lin. Support Vector Machine Solvers. Available at <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.64.4200&rep=rep1&type=pdf>
- [21] Support Vector Machine, http://pages.cs.wisc.edu/~jerryzhu/cs540/handouts/hearst_98-VMtutorial.pdf
- [22] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin, “A Practical Guide to Support Vector Classification”
- [23] West M, Blanchette C, Dressman H, et al. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc Natl Acad Sci USA*. 2001;98 (20):11462–11467.
- [24] Jungermann AH. Entropy and the Shelf Model: a quantum physical approach to a physical property. *J Chem Educ*. 2006;83(11):1686–1694.
- [25] Granzow M, Berrar D, Dubitzky W, et al. Tumor classification by gene expression profiling: comparison and validation of five clustering methods. *Acm Sigbio Newsletter*. 2001;21:16–22. [2] Jiang D, Tang C, Zhang A. Cluster analysis for gene expression data: a survey. *IEEE Comput Soc*. 2004;11:1370–1386
- [26] Bollschweiler EH, Monig SP, Hensler K, et al. 2004. Artificial neural network for prediction of lymph node metastases in gastric cancer: a phase II diagnostic study. *Ann Surg Oncol*, 11:506-11.
- [27] Bottaci L, Drew PJ, Hartley JE, et al. 1997. Artificial neural networks applied to outcome prediction for colorectal cancer patients in separate institutions. *Lancet*, 350:469-72.
- [28] Bryce TJ, Dewhirst MW, Floyd CE Jr, et al. 1998. Artificial neural network model of survival in patients treated with irradiation with and without concurrent chemotherapy for advanced carcinoma of the head and neck. *Int J Radiat Oncol Biol Phys*, 41:239-45.
- [29] Burke HB, Bostwick DG, Meiers I, et al. 2005. Prostate cancer outcome: epidemiology and biostatistics. *Anal Quant Cytol Histol*, 27:211-7. *Burke HB, Goodman PH, Rosen DB, et al. 1997. Artificial neural networks improve the accuracy of cancer survival prediction. *Cancer*, 79:857-62.
- [30] Catto JW, Linkens DA, Abbod MF, et al. 2003. Artificial intelligence in predicting bladder cancer outcome: a comparison of neurofuzzy modeling and artificial neural networks. *Clin Cancer Res*, 9:4172-7.
- [31] Dimitoglou, G., Adams, J. A., & Jim, C. M. (2012). Comparison of the C4.5 and a naive Bayes classifier for the prediction of lung cancer survivability. *Journal of Computing*, 4(8), 1–9.
- [32] Dinh, A., Miertschin, S., Young, A., & Mohanty, S. D. (2019). A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. *BMC Medical Informatics and Decision Making*, 19(211), 1–15.
- [33] Munsell, B. C., Wee, C. Y., Keller, S. S., Weber, B., Elger, C., da Silva, L. A. T., ... Bonilha, L. (2015). Evaluation of machine learning algorithms for treatment outcome prediction in patients with epilepsy based on structural connectome data. *NeuroImage*, 118, 219–230.
- [34] Nilashi, M., binIbrahim, O., Ahmadi, H., & Shahmoradi, L. (2017). An analytical method for diseases prediction using machine learning techniques. *Computers & Chemical Engineering*, 106, 212–223.
- [35] Okada, H., Hontsu, S., Miura, S., Asakawa, I., Tamamoto, T., Katayama, E., ... Hasegawa, M. (2012). Changes of tumor size and tumor contrast enhancement during radiotherapy for Non-small-cell lung cancer May Be suggestive of treatment response. *Journal of Radiation Research*, 53(2), 326–332.
- [36] Oztekin, A., Delen, D., & (James)Kong, Z. (2009). Predicting the graft survival for heart–lung transplantation patients: An integrated data mining methodology. *International Journal of Medical Informatics*, 78(12), e84–e96. Park, S., Lee, S. J., Weiss, E., & Motai, Y. (2016).