# Applying CNN on Lung Images for Screening Initial Cancer Stages

Susmitha Valli Gogula
*Computer Science and Engineering*
*Gokaraju Rangaraju Institute of Engineering and Technology*
Hyderabad, India
susmitagv@gmail.com

Srinath Goud Komiri
*Computer Science and Engineering*
*Gokaraju Rangaraju Institute of Engineering and Technology*
Hyderabad, India
komirisrinathgoud99@gmail.com

Vishnu Vardhan Arivilli
*Computer Science and Engineering*
*Gokaraju Rangaraju Institute of Engineering and Technology*
Hyderabad, India
arivilli688@gmail.com

Pandari Gulapally
*Computer Science and Engineering*
*Gokaraju Rangaraju Institute of Engineering and Technology*
Hyderabad, India
pandarigulapally2000@gmail.com

Kowshik Kodumuri
*Computer Science and Engineering*
*Gokaraju Rangaraju Institute of Engineering and Technology*
Hyderabad, India
kowshikkodumuri@gmail.com

2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC) | 978-1-6654-9710-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICAAIC53929.2022.9793298

***Abstract*— Cancer is the most common disease nowadays, in particular the lung cancer is often diagnosed in many individuals. There are various factors that contribute to cancer in humans, but among them tobacco smoking remains as a key contributor. Smoking is a primary cause, but many other variables, such as second-hand smoke, industrial pollutants, asbestos exposure, and so on, can also cause cancer. These all need to be filtered by lungs, as lungs always need to be working, unlike other organs in the body lungs does not have any rest so it gets effected early than all other body organs, in such cases it has to be examined carefully and clearly for many times to conclude whether it is affected with cancer. This effort is made to make such simple support for doctors in designing a CAD system for identifying the presence of tumor in lungs, this model has high accuracy rate in identifying the problem. CNN is a reliable algorithm for finding such minute problems in CT Scan image of lung to confirm the disease. Lung CT images were used in this study. Training accuracy of our model is 96.11% and the validation accuracy is 97.8%.**

***Keywords*— *Computer Aided Design, Convolutional Neural Networks, Computed Tomography, Deep Learning, Lung Cancer***

## I. INTRODUCTION

Lung cancer is a type of cancer that starts in the lungs and expands to various parts in the body. Cancer in the lungs is the most widely recognized sort of disease that kills people. The assumption is that genetic factors must put certain individuals at higher risk for cellular breakdown in the lungs after openness to cancer-causing agents [6]. Lung cancer was diagnosed in an approximated 171,600 people in the United States in 1999 (94,000 men and 77,600 women), with 158,900 people dying as a result of the disease. As a safety measure, the United States Preventive Services Task Force (USPSTF) indicates that high-threat adults be checked yearly with low-dose computed tomography. (CT) [23]. For the reasons stated above, it is necessary to deploy a CAD system to assist clinicians in identifying lung cancer as early as possible, not only recognising the nodule but doing so with high accuracy. Our aim is to recognize the presence of cellular breakdown in the lungs in understanding CT images of lungs with and without early phase cellular breakdown in the lungs, using a binary classification issue. To create an accurate classifier, this research work attempts to leverage different approaches from computer vision and deep learning, specifically convolutional neural networks. This research study has used a dataset from Kaggle and constructed a CNN model, trained for the purpose of Lung Cancer detection.

## II. BACKGROUND

In the lung cancer diagnosis, computed tomography (CT) is needed to spot the pulmonary nodules. To detect and categorise pulmonary nodules in clinical CT scans, we need to employ a well-trained deep learning system, as deep learning algorithms have recently been recognised as a promising tool in the medical field.[4]

This study was designed to aid doctors in making decisions regarding a patient's health and increase informed patient consent by providing a thorough grasp of the risks involved in treatment procedures based on the patient's condition. By gathering information about the patient's state, we can also save some expensive resources that aren't required for the patient. Despite ongoing forward leaps in analytic strategies, unobtrusive changes, and theoretical therapies, cellular breakdown in the lungs patient results stay poor; subsequently, a more profound comprehension of hazard variables might affect local area level preventive drives [1].

Convolutional neural network (CNN) was the primary deep learning technique to acquire widespread attention for their superior performance in AI applications [16]. Several medications developed as a result of these research are now approved for the treatment of certain types of lung cancer. Current lung cancer biology research using cell lines, tumour samples, and animal models, along with knowledge of the lung cancer genome, will bring about a superior comprehension of the illness and new remedial options for patients [21].

Victor [10] employed a deep learning model and achieved an accuracy of 88.4%. Jan et al. [18] A morphological and circular filter-based lung segmentation approach was proposed. Later, they have used CNN approach and got an accuracy of 84.62%. Lyu. [11] developed a Multi-Level

CNN to detect the cellular breakdown in the lungs and got an accuracy of 84.81 percent.

## III. Literature Review

Victor [10] employed a deep learning model and achieved an accuracy of 88.4%. Jan et al. [18] A morphological and circular filter-based lung segmentation approach was proposed. Later, they have used CNN approach and got an accuracy of 84.62%. Lyu. [11] developed a Multi-Level CNN to detect the cellular breakdown in the lungs and got an accuracy of 84.81 percent.

The basic SVM is a binary linear non-probabilistic classifier that accepts a stream of input data and estimates which of the two classes each input belongs to. An SVM technique constructs a model that allocates fresh instances to one of two categories based on a pile of training example data that has been labelled as belonging to one of two categories.[5] Radhika P.R., Veena G.

Suren Makajua, P.W.C. Prasad, Abeer Alsadoona, A. K. Singhb, A. Elchouemic. "The proposed system employs the linear classifier technique (SVM**).** In the instance of the SVMclassifier, only two features are chosen for classification out of nine at a time, resulting in a benign or malignant outcome[14].

W. Rahane, H. Dalvi, Y. Magar, A. Kalane, and S.Jondhale [8] proposed using image processing and ML (Support Vector Machine) for lung cancer recognition on CT images. Highlights like area, parameter, and eccentricity from the portioned picture locale of interest are given as input to the support vector machine (SVM) model.

Mr. Vishal Patil, Dr. Aditya Gupta proposed using image processing and K-Means clustering using two approaches i.e., using centroid and using Euclidean distance. Also stated the weakness of the algorithm as, in K-Means clustering the result depends on the value K [1].

The Deep Learning algorithm, according to Issa Ali, Gregory R. Hart, and Gowthaman Gunabushanam, receives a CT picture as input and perceives it as a collection of states, resulting in a categorization of whether a nodule is there or not. The researchers looked at a total of 800 CT images and discovered that 590 people had one or more nodules, whereas 298 had none. The target of their work is to create and approve a model in view of profound counterfeit neural organizations for early identification of lung knobs in thoracic CT pictures [12].

Eali Stephen Neal Joshua1, Midhun Chakravarthy, Debnath Bhattacharyya clarified the disadvantages of many AI calculations like K Nearest, Back propagation network, Decision Fusion, Naïve Bayes, Support vector machine, Random forest and so on, After a deliberate writing review, they discovered that a few classifiers have less precision and some are high exactness yet not came to closer of 100 percent. After a broad review, they observed that gathering classifier was beat when contrasted and the other AI calculations [3].

To detect lung cancer using CT images, Qing Wu devised an unique neural-network based approach known as the entropy degradation method (EDM). His algorithms have a 77.8 percent accuracy rate [13].

The CNN separates picture highlights to produce progressively complex portrayals, as indicated by Stanley Cohen MD. It employs feature extraction filters on the source image and some hidden layers to progress from low to high level feature maps [4].

## IV. Methodology

Our proposed system contains the data collections, data formatting, model training, testing, and prediction utilising K-Means, KNN, SVM and CNN detailed in the following sections.

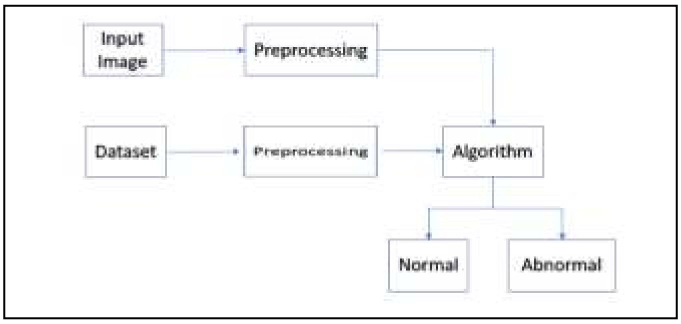

Fig. 1. System Architecture

### A. Data Acquisition

The lung CT images are obtained from Kaggle website. The images are classified into two types: Normal and Abnormal. The dataset contains 138 images in total, where 58 are abnormal and 80 are normal.

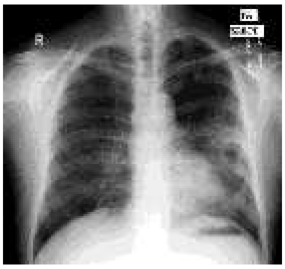

Fig. 2. Abnormal

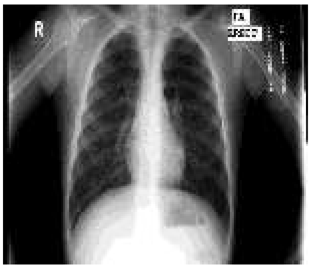

Fig. 3. Normal

### B. Data Formatting

The images were resized to maintain a uniform aspect ratio of one with (400, 400) pixel size for the CNN operation. All the pixel values for the images were converted in the range of (0,1) to make the convergence faster.

### C. Model Training, Testing and Prediction

Images are split in a ratio of 80:20 for training and testing purposes.

*a) K-Means Algorithm:* Clustering with k-means is a basic but effective method. Traditionally, k data points are picked at random as cluster centres, or centroids, from a given dataset, and all training examples are plotted and added to the closest cluster. After all instances have been added to clusters, the centroids, which reflect the mean of each cluster's instances, are recalculated, and these recalculated centroids become the new cluster centres.

Step 1: Initialize 2 random clusters with centroids as A and B where A is Normal type of image and B is Abnormal type of image.

Step 2: The Algorithm assigns each data point P to it's nearest cluster by calculating the Euclidean distance, the Euclidean distance is calculated as follows:

$$Dis= L\ (A, P) \tag{1}$$

Where L is the Euclidean calculation function it can be calculated as:

$$L^2 = \Sigma \quad (x_i\text{-}y_i)^2 \tag{2}$$

Where $x_i$, $y_i$ are Euclidean vectors, starting from the origin of the space.

Then it repeats the process for all the image data points and classifies the images into 2 categories

### *b) KNN Algorithm:*

In KNN algorithm the classification is done by considering the majority vote of neighbours. The following Fig 3 shows the flow of KNN classification algorithm. The distance is calculated in the same way which we have used in case of K-Means algorithm i.e by Euclidean distance.

Decide on the number of neighbors (K), then determine the Euclidean distance between K neighbors. Using the obtained Euclidean distance, find the K closest neighbors. Count the number of data points in each category among these k neighbors. Assign the new data points to the category with the greatest number of neighbors.

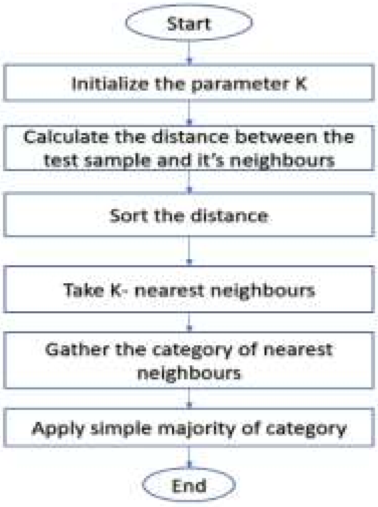
Start
Initialize the parameter K
Calculate the distance between the test sample and it's neighbours
Sort the distance
Take K- nearest neighbours
Gather the category of nearest neighbours
Apply simple majority of category
End

Fig. 4. KNN classification algorithm

### *c) SVM Algorithm:*

SVM is an administered AI strategy that can be utilized to tackle, order and relapse issues. The support vector machine's calculation will probably find a hyperplane in a N-layered space (N - the quantity of elements) that recognizes important elements. We may obtain a large number of hyperplanes, but we select the one with the greatest distance from the support vectors (the nearest data items from each group). The greater the edge distance, the easier it is to classify the bunch of input items. The SVM classifier is fitted to the prepared information.

### *d) CNN Algorithm:*

The Convolutional Neural Network (CNNs) for image categorization and recognition was built using a group of layers. Convolutional layers with kernel filters, max/avg pooling, and fully linked layers were used to process training and testing pictures. To classify the provided item, the softmax function was used. For this job, a neural network with three hidden layers, one input layer, and one fully connected layer was used. The input layer receives images with a resolution of (400, 400). In each convolutional layer, a kernel matrix of (3, 3) was used to extract the features and we have used an activation function, ReLU, that returns only the important features by keeping the positive values as same and reducing negative values to 0. To decrease the computing parameters in the following convolution layer, a maximum pooling size of (2, 2) was implemented. The model was given a dropout value of 0.1. The class probabilities for final output classes were calculated using a dense value of two and the softmax activation function. The learning rates for various parameters were calculated using an adaptive moment estimation (Adam) optimizer. The disparity between the anticipated output and the labelled output for the given input is calculated using a loss function; categorical cross-entropy (CE) was utilized for this challenge.

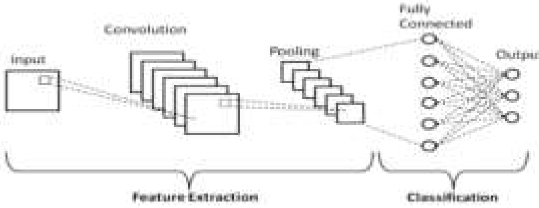

Fig. 5. CNN Architecture

*Pooling Layer:* In general, pooling indicates a small part of the input, so we take a little part of the input and try to take the average value (average pooling) or take the maximum value (max pooling), so when we pool an image, we are obtaining a summarized value across all of the data there.

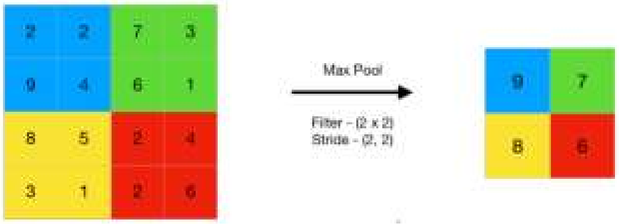

Fig. 6. Max pooling

*ReLU Activation Function:* The ReLU activation function can be calculated as Max(0,z), where z is any value in the input vector.

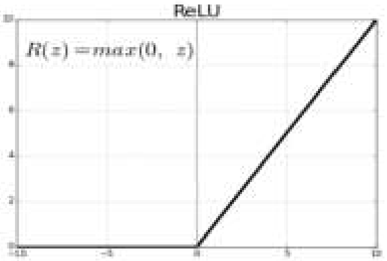

Fig. 7. ReLU Activation Function

*Softmax Activation Function:*

The softmax function is most commonly used as an activation function in a neural network model in applied machine learning. The network is set up to produce N values, one for each classification task class, and the softmax function is used to normalize the outputs, turning them from weighted sum values to probabilities that total to one. Each number in the softmax function's output is interpreted as the likelihood of belonging to each class.
Softmax can be calculated as:

$$S(x_i)= \frac{}{\Sigma_j^n} \tag{3}$$

Where x is an input vector to a softmax function S. It consists of n elements for n classes.
$x_i$ is the i-th element of the input vector. It can take any value between -inf and +inf.
The normalization term in the denominator ensures that the values of the output vectors sums to 1 for i-th class and each of them is in the range 0 and 1 which makes up a valid probability distribution.
n is the number of classes.

TABLE I. CNN ARCHITECTURE
(DROPOUT WITH 0.2, ADAM OPTIM, LEARNING RATE = 0.001)

| Layer | Params | Activation | Output |
|---|---|---|---|
| ***Input*** | | | 256x256x1 |
| Conv1 | 3x3x32 | ReLU | 256x256x32 |
| MaxPool | 2x2, Stride 2 | | 128x128x32 |
| Conv2 | 3x3x32 | ReLU | 128x128x32 |
| MaxPool | 2x2, Stride 2 | | 64x64x32 |
| Dense | | | 256 |
| Dense | | | 2 |

## V. RESULTS

In all algorithms training and testing images were split in the ratio of 80:20, the K-Means Clustering model achieved a validation accuracy of 38.33%. KNN algorithm achieved an accuracy of 59.2%. The SVM algorithm achieved an accuracy of 50.1%. In CNN the images were trained for 10 epochs with batch size 32 in each epoch. Model acquired 96.11% training accuracy and 97.8% validation accuracy in the final epoch.

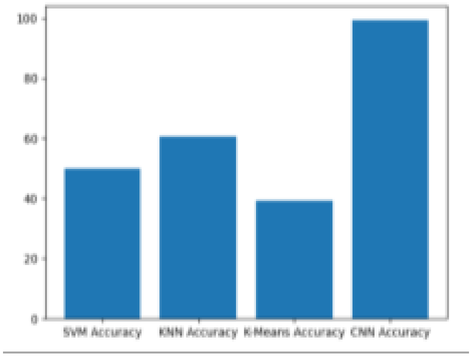

Fig. 8. Accuracy Graph

The above graph illustrates the performance of SVM, KNN, K-Means and CNN algorithms on our dataset.

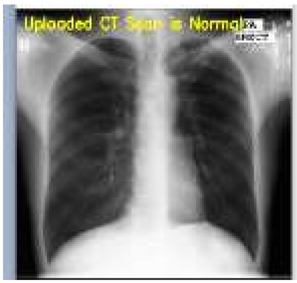

Fig. 9. Normal

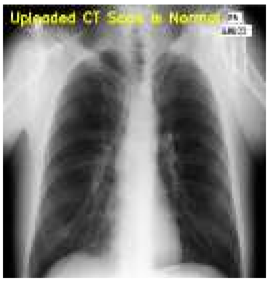

Fig. 10. Normal

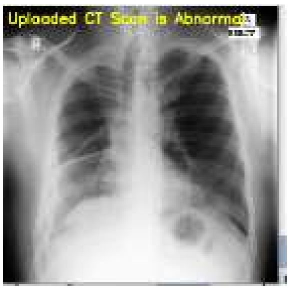

Fig. 11. Abnormal

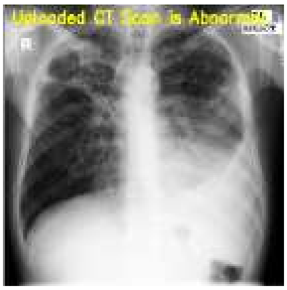

Fig. 12. Abnormal

The above Fig. 9, Fig. 10 are the results of a CT image of lungs of a normal patient. The Fig. 11, Fig. 12 are the results of lungs affected with cancer.
The accuracy and precision of the developed CNN model are calculated as below.

$$\text{accuracy} = \frac{TN)}{} \tag{4}$$

$$\text{precision} = \frac{}{FP)} \tag{5}$$

Where, TP is True Positive, TN is True Negative, FP is False Positive, FN is False Negative.
The following table is the confusion matrix obtained from the developed CNN model. Accuracy is 97.8, Precision is 96.5.

| N=138 | **Predicted** | |
|---|---|---|
| **Actual** | ***Abnormal*** | ***Normal*** |
| Abnormal | 56 | 2 |
| Normal | 1 | 79 |

TABLE II. CONFUSION MATRIX

## CONCLUSION AND FUTURE SCOPE

In this study we have examined some machine learning models like K-Means, KNN, SVM and CNN in detecting the lung cancer and observed that the CNN model is providing us the better accuracy among all the models. We employed a CNN classifier to evaluate whether a CT image of the lung was cancerous or not in this study. Utilising the Kaggle dataset, we conducted a detailed experiment. Our detection accuracy is 97.8 percent, which is higher than that of previous approaches. In the future, we'll run the studies on a larger dataset and include new variables like nodule size, texture, and position to improve the results even more. We'll also aim to leverage deep CNN approaches to improve accuracy, and we'll apply our technology to other forms of cancer diagnosis.

## REFERENCES

[1] Mr. Vishal Patil, Dr. Aditya Gupta "Lung Cancer Detection Using Image Processing" JETIR March 2021s.

[2] Yang Song, Weidong Cai, in Computer Vision for Microscopy Image Analysis, https://www.sciencedirect.com/science/article/pii/B9780128149720000047, 2021.

[3] Eali Stephen Neal Joshual, Midhun Chakkravarthy, Debnath Bhattacharyya "An Extensive Review on Lung Cancer Detection Using Machine Learning Techniques: A Systematic Study" Revue d'Intelligence Artificielle, researchgate, 7 May 2020.

[4] Stanley Cohen MD "Knowledge Discovery in Big Data from Astronomy and Earth Observation" Astrogeoinformaticshttps://www.sciencedirect.com/book/9780128191545/knowledge-discovery-in-big-data-from-astronomy-and-earth-observation 2020.

[5] Radhika P.R, Rakhi A.S Nair, Veena G "A Comparative study of lung cancer detection using machine learning algorithms", 2019.

[6] Steven E. Weinberger MD, MACP, FRCP, Jess Mandel MD, FACP, in Principles of Pulmonary Medicine (Seventh Edition), 2019

[7] Sajja Tulasi Krishna, Retz Mahima Devarapalli, Hemantha kumar Kalluri Lung Cancer Detection Based on CT Scan Images by Using Deep Transfer Learning

[8] Article in Traitement du Signal · October 2019 https://www.researchgate.net/publication/336879291

[9] Wasudeo Rahane, Himali Dalvi "Lung Cancer Detection Using image Processing and Machine Learning", IEEE 29 November 2018

[10] H. Peschl, D. Han, P. Van Ooijen, M. Oudkerk, Lung Cancer prediction using deep learning software: validation on independent multi-centre data, J. Thorac. Oncol. 13 (2018) S428,

https://doi.org/10.1016/j.jtho.2018.08.489.

[11] Victor, R., Peixoto, S., Pires, S., Silva, P., Pedrosa, P., Filho, R. (2018). Lung nodule classification via deep transfer learning in CT lung images. In 2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS), Karlstad, Sweden. https://doi.org/10.1109/CBMS.2018.00050

[12] Lyu, J., Ling, S.H. (2018). Using multi-level convolutional neural network for classification of lung nodules on CT images. In 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, USA, pp.686-689. https://doi.org/10.1109/EMBC.2018. 8512376

[13] Issa Ali, Gregory R.Hart, Gowthaman Gunabushanam "Lung Nodule Detection via Deep Reinforcement Learning" PMID: 29713615 2018 Apr 16.

[14] Qing Wu "Small-Cell Lung Cancer Detection Using a Supervised Machine Learning Algorithm" International Symposium on Computer Science and Intelligent Controls 2017

[15] Suren Makajua, P.W.C. Prasad, Abeer Alsadoona , A. K. Singhb , A. Elchouemic" Lung Cancer Detection using CT Scan Images " ICSCC 2017, 7-8 December 2017.

[16] F. Ciompi, K. Chung, S.J. Van Riel, A.A.A. Setio, P.K. Gerke, C. Jacobs, E. Th Scholten, C. Schaefer-Prokop, M.M.W. Wille, A. Marchiano, ` U. Pastorino, M. Prokop, B. Van Ginneken, Towards automatic pulmonary nodule management in lung cancer screening with deep learning, Sci. Rep. 7 (2017), https://doi.org/ 10.1038/srep46479

[17] Yi- long Wu, Chao Zhang "Toward an Expert Level of Lung Cancer Detection and Classification Using a Deep Convolutional Neural Network" PMID: 30996009 2019 Apr17.

[18] Yaron B.Gesthalter, EhabBillatos, HasmeenaKathuria "Genomic and Precision Medicine (Third Edition)" science direct 2017

[19] Jin, X., Zhang, Y., Jin, Q. (2016). Pulmonary nodule detection based on CT images using Convolution neural network. 2016 9th International Symposium on Computational Intelligence and Design (ISCID), Hangzhou,China. https://doi.org/10.1109/ISCID.2016.1053

[20] Centers for Disease Control and Prevention, "Lung cancer statistics." https://www.cdc.gov/cancer/ lung/statistics/, 2016.

[21] Emre Dandal, Murat Cakiroglu, Ziya Eksi, Murat Ozkan "Artificial neural network-based classification system for lung nodules on computed tomography scans" 6th International Conference of Soft Computing and Pattern Recognition At: Tunisia, August 2014.

[22] K.Politi, C.S.Dela Cruz, R.Homer "Pathobiology of Human Disease" A Dynamic Encyclopedia of Disease Mechanisms science direct 2014

[23] Disha Sharma, Gagandeep Jindal "Identifying lung cancer using Image Processing Techniques"International Conference on Computational Techniques and Artificial Intelligence (ICCTAI'2011)

[24] Tai Lahans "Integrating Conventional and Chinese Medicine in cancer care" A clinical guide 2007

[25] Kenji Suzuki,Samuel G. Armato III,Feng Li,Shusuke Sone,Kunio Doi "Massive training artificial neural network (MTANN) for reduction of false positives in computerized detection of lung nodules in low-dose computed tomography" PMID: 12906178, ncbi JULY 2003.

[26] SUSAN T. MAYNE, in Nutrition in the Prevention and Treatment of Disease, https://www.sciencedirect.com/topics/social-sciences/lung-cancer 2001