IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol. 9, Issue.5, May 2021 E-ISSN: 2347-2693

Image classification Method in detecting Lungs Cancer using CT images: A Review

Astha Pathak[1*], Avinash Dhole[2]

[1,2]Dept. of CSE, RITEE Raipur, C.G. India

[*]Corresponding Author: asthapathak1806@gmail.com

DOI: https://doi.org/10.26438/ijcse/v9i5.3742 | Available online at: www.ijcseonline.org

Received: 07/May/2021, Accepted: 17/May/2021, Published: 31/May/2021

Abstract— A tumour is an irregular mass of cells and it can either be benign (non-cancerous) or malignant (cancerous). Disease alludes to cells that outgrow control and attack different tissues. One of the reasons for malignancy passing in person is Lung Cancer. Clinical therapy with drugs intended to target lungs disease cell to diminish the spread all through the body may likewise conceivable yet before this it is must to perceive the malignant growth at the beginning phase. Physically disease recognizable proof is tad of tedious so that with the progression of innovation, Several Computer Aided Diagnosis (CAD) frameworks are created for distinguishing cellular breakdown in the lungs in its beginning phase. In this paper inclination in detail literature survey on various techniques that have been used in feature extraction and classification with its obtain accuracy.

Keywords—CAD, SIFT, SVM, ANN

I. INTRODUCTION

In today's lifestyle cancer become the one of the common disease to occur but their treatment become complicated due to its late detection. Early detection of cancer helps to prevent its spread all over the body. In the field of computing that is geared toward guaranteeing the success of the machine in such issues, is grow day by day. One in all the large step with in the history of computing is that the finding of artificial neural networks. It is satisfying for several years. They need not been able to meet the strain of the business field particularly with advancing technology. At now, multi-layered structure known as "Deep Learning" are emerged[1]. Image classification may be a problem of computer vision that deals with tones of basic information such as agriculture, meteorology and safety. The human brain can easily classify images but for the computer it is not easy if the image contain noise. Different methods are develop to perform the classification operation. General classification procedure are often divided into two broad categories as Supervised classification supported the tactic used and unsupervised classification [21]. Several Computer Aided Diagnosis (CAD) systems are developed for detecting carcinoma in its early stage. A CAD system may consists of several steps in determining carcinoma. They are as follows (1) Preprocessing or lung Segmentation (2) Nodule detection (3) Nodule segmentation (4) Feature Extraction (5) Classification. Fig. 1. show the steps of the CAD system in detecting Lung cancer.

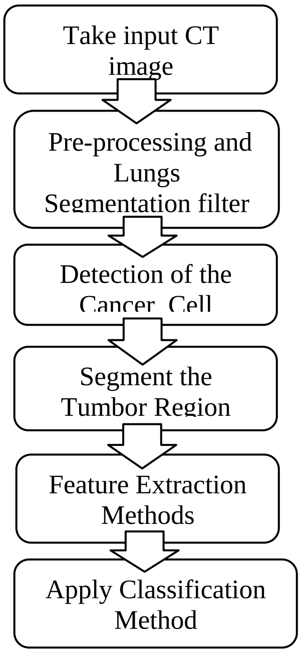
Take input CT image
Pre-processing and Lungs Segmentation filter
Detection of the Cancer Cell
Segment the Tumbor Region
Feature Extraction Methods
Apply Classification Method

Fig 1: Traditional CAD Process

Preprocessing:- It involves removal of noise from the image, improving the quality of the image by using filters and separating the lung region from the CT slice. Nodule detection: - In this stage, the candidate nodules are identified by enhancing the suspected areas and suppressing the other structures like blood vessels. This step reduces the search space for nodule detection.

Features Extraction:- In this step is identification of the nodule as benign or malignant nodule has been done . Features include size, shape and appearance of the nodule or growth rate of the nodule. Nodules of size but 1 cm are likely to be benign and greater than 1 cm are malignant nodules. The shape of the benign nodule is circular, whereas for malignant nodule it is Lobulated, speculated, ragged and halo. Benign nodules tend to have even surface but malignant have uneven surface. With reference to rate of growth malignant nodules grow faster comparatively with benign nodules. Classification:-Classification is a supervised learning problem which defines a set of target classes, and trains a model to recognize them using labeled example photos.

In this paper discussion about different image features extraction and image classification algorithm with comparison in their obtained accuracy has been done. Rest of the paper is organized as follows, Section II contain the related work of Literature of image classification and feature extraction algorithm, Section III contain the some methodology of image classification algorithm , Section IV describes results and discussion.

II. Related Work

Emine CENGIL and Ahmet CINER [1] proposed a system to lungs cancer identification and use Tensor flow liberary for cancer diagnosis and 3D CNN architecture for system training. While classifying the image they used CT image of SPIE-AAPM-LungX database and obtain an accuracy of 70%.

T. N. Shewaye et al. [2] proposed an automated system to lungs classification by using data set of LICD-IDRI CT data set. They applied a combination of geometric and histogram lungs nodules image features and different linear and non linear discriminate classifier .The system can classify 82% of malignant and 93% of benign nodules on unseen test data. A. M. Suzan et al.[3] classification of lungs cancer is based on a code book generated by using bag of feature algorithm by using SPIE database. They used scale invariant features transform (SIFT) for feature extraction and this coefficient are quantized using bag of features into predefine code book and obtain an accuracy of 95% for code book size 500. R.Anirudh et al.[4] makes use of 3D CNN and trained using weak level label information with sensitivity of 80% for 10 false positive per scan.

Q.Song et al [5] makes use of three different neural network architecture on the same basis and found CNN as the best precious with an accuracy of 84.15% by using LIDC-IDRI dataset.

S. M Salaken et al [6] proposed a deep automated classifier mechanisms for the low population dataset and the performance improvement is satistically significant with the accuracy of 80%.

M.F Sergece et al. [7] proposed a deep learning architecture with low variance in medical binary classification tasks for learning high-level image representation to achieve high accuracy. They evaluate their model on Kaggle Data Science Bowl 2017 [KDSB17] data set and compare the result with the related work proposed in the kaggle competition and achieved a sensitivity of 0.87, Specify of 0.991 and log loss of 0.20.

E.Cengil et al. [8] proposed its image classification process by using CNN. They used a caffe library in Kaggle dataset by GPU technology and evaluated the classification with an accuracy rate of 83%.

D. Jayaraj et al. [9] proposed a new automated model to identified lungs cancer by using LIDC dataset. They used median and Gaussian filter in preprocessing and segment the image by using watershed segmentation algorithm. The classification of image will be done by using RF classifier and obtain a maximum accuracy of 89.90%. Nidhi S. Nadkarni. et al.[10] proposed a system for lungs cancer identification by using TCIA dataset. They applied median and contrast adjustment filter in image pre-processing and then SVM algorithm in the classification system to detect whether the patient is cancerous or not.

Kuntal Pal et.al[11] shown the important of pre-processing technique in three variation of the CNN by using CIFARIO dataset. They analyze the result by using three different pre-processing technique and then obtain an accuracy of 55-58% in mean normalization technique, 63-66% in standardization and 64-68% in ZCA technique.

Michael Blot et.al.[12] presents the new architecture of CNN ,Max-Min CNN on two different database namely MNIST and CIFAR-10 and identified that the strategy reaches very good performance on CIFAR-10 dataset with an accuracy of 78.62% and MNIST with an accuracy of 99.34%.

Travis Williams et.al.[13] proposed method by using MNIST and CIFAR-10 dataset and preprocess the data in the wavelet domain by separating the image into two different sub bands and learn features by varying the frequency in low to high and obtain that CNN-WAV4 method achieve the highest classification accuracy of 85%. K.Gopi et.al.[14] proposed a system to classify lungs tumor recognized area and detection of tumbor by using LIDC dataset of lungs image. The method preprocessing the input image by using median and wiener filter, segmentation can be done by using EK algorithm, features extraction by using GICM and then finally classified by the SVM algorithm. This method provides an accuracy of 92.46%.

S Khobragade et.al[15] proposed automated system to detect lungs disease like TB, Pneumonia and lungs cancer by using chest radiography database of 80 patients from Sansoon hospital Pune. They applied histogram equalization method to enhance the image, intensity and discontinuity based method to obtaining the region of interest and image classification by Artificial Neural Network. The automated system obtained an accuracy of 92%.

Sayali Satish Kanitkar et.al[16] thresholding and marker control watershed transform is applied in segmentation of CT scan image of the lungs. For the system database are collected from the hospital and while comparing the both segmentation method they obtain a best accuracy of 100% in marker control watershed transform segmentation technique. Sheenam Rattan et.al[17] detection of cancer in early stage can be successfully done by applied watershed transform and for the best solution BAT solution has been used in segmentation of CT scan image of lungs and in classification ANN has been applied. The accuracy obtain by the system is 98.5%.

Anam Tariq et.al[18] detections of lungs nodules by segmentation and classification of CT Scan image has been done by using the Neuro fuzzy based classifier by Appling post processing phase ROI was extracted and then successfully obtain an accuracy of 95%.

Nooshin Hadavi et.al[19] CAD system was designed which aim to fast and accurately detects lungs cancer. The system used database of 60 CT scan image in which 70% were used in training and 30% were used in testing the proposed technique. During segmentation section region growing algorithm has been applied on the CT image and then extract the ROI, then by using band pass filter, nodule of cancer has been identified. The proposed approach was success to detect 90.9% of cancer cases.

S.K. Vijai Anand[20] proposed system which can detect the lungs cancer in 3 minutes. The system used CT scan image of patients and applied GICM for features extraction and then back propagation network to classified the image in cancerous or not. The proposed system obtained an accuracy of 86.3%.

Emine Cengil et.al[21]proposed system which can successfully distinguishes 1.2 million images with 1000 categories. This was performed in the Caffe liberary by using the Alexnet model and constructed by using CNN.Caltech-101 dataset was used to see the performance of the model and then obtain a high accuracy.

In 2020 Rahul Meena, Vighnesh Menon, Vivek Solavande[22] use CNN architecture for image Pre-Processing and their model can detect many disease related to respiratory system and achieved an accuracy of 92.30%.
In 2020 Zarli Cho1 , Khin Myo Kyi , Kyi Thar Oo[23] presents the concept of image feature extraction with Alexnet CNN model by using data base of United States Geological Survey (USGS) and obtain an accuracy of more than 66%.

Table1 shows the different classification methods with its limitation based on the above literature survey.

Ref No.	Author Name	Classification Method	Dataset	Limitation
[1]	E CENGIL et.al	Tensorflow library	SPIE-AAPM-LungX database	Focus only on black and white image and architecture need to be improved
[2]	T. N. Shewaye et al.	linear and nonlinear discriminate classifier	LICD-IDRI CT data set.	Absence of Texture image feature in input dataset.
[3]	A. M. Suzan et al.	Bag of features	SPIE-Dataset	Code book size is small and not only consider the cell area for classification
[6]	S. M Salaken et al	Deep NN	UCI machine learning repository dataset	In terms of accuracy
[8]	E.Cengil et al.	Caffe Library	Kaggle Dataset	In terms of speed and accuracy
[9]	D. Jayaraj et al.	RF classifier	LIDC dataset	Only focus on black and white images and accuracy may also increase.
[10]	Nidhi S. Nadkarni. et al.	SVM algorithm	TCIA dataset	Accuracy may increase by using different classification methods.
[13]	Travis Williams et.al.	CNN-WAV4	MNIST ,CIFAR-10	Size of Dataset is small
[14]	K.Gopi et.al	Support Vector Machine	LIDC	Segmentation method required to improve to increase the accuracy rate.
[15]	S Khobragade et.al	Artificial Neural Network	Private Dataset	System is not robust and unable to diagnosis the lungs cancer accurately.
[17]	Sheenam Rattan et.al	Artificial Neural Network	Private Dataset	Processing of an image is done on only grey scale images not in colour image.
[18]	Anam Tariq et.al	Neuro fuzzy classifier	Private Dataset	Work with only 100 images.
[19]	Nooshin Hadavi et.al	Cellular Automata	Private Dataset	In terms of accuracy

III. METHODOLOGY

In CAD Process there are two main important techniques which play a very important role in cancer cell detection. They are as follows (1) Feature Extraction Techniques (2) Classification Techniques.

Feature Extraction:- Feature extraction may be a process of dimensionality reduction by which an initial set of data is reduced to more manageable groups for processing. A characteristic of those large data sets may be a sizable amount of variables that need tons of computing resources to process. Feature extraction is that the name for methods that selects and /or combines variables into features, effectively reducing the quantity of knowledge that has got to be processed, while still accurately and completely describing the original data set. Some of the features extraction methods are:-

GLCM:- Gray level co-occurrence matrix (GLCM) is a popular texture-based feature extraction method. The GLCM determines the textural relationship between pixels by performing an operation consistent with the second-order statistics within the images. Usually two pixels are used for this operation. The GLCM determines the frequency of combinations of those pixel brightness values determined. That is, it represents the frequency formation of the pixel pairs . The GLCM properties of a picture are expressed as a matrix with an equivalent number of rows and columns because the gray values within the image. The elements of this matrix depend on the frequency of the two specified pixels. Both pixel pairs can vary depending on their neighbourhood. These matrix elements contain the second-order statistical probability values depending on the gray value of the rows and columns. If the intensity values are wide, the transient matrix is quite large. This creates a time-consuming process load . The GLCM features utilized in this study are as follows; autocorrelation, contrast, correlation, cluster prominence, cluster shade, dissimilarity, energy, entropy, homogeneity, maximum probability, sum of squares (variance), sum average, sum variance, sum entropy, difference variance, difference variance, difference entropy, system of measurement of correlation, inverse difference normalized, inverse difference moment normalized. A GLCM feature matrix is generated which can successfully represent a picture with fewer parameters using these properties [22].

SIFT:- SIFT(scale-invariant feature transform) is quite an involved algorithm. There are mainly four steps involved within the SIFT algorithm.

- Scale-space peak selection: Potential location for locating features.
- Localization: Accurately locating the feature key points.
- Orientation Assignment: Assigning orientation to key points.
- Key point descriptor: Describing the key points as a high dimensional vector.
- Key point Matching: Key points between two images are matched by identifying their nearest neighbours.

Band pass filter: - Band pass filters are a mixture of both low pass and high pass filters. They attenuate all frequencies smaller than a frequency D0 and better than a frequency D1, while the frequencies between the 2 cut-offs remain within the resulting output image. We obtain the filter function of a band pass by multiplying the filter functions of a low pass and of a high pass within the frequency domain, where the cut-off frequency of the low pass is higher than that of the high pass. Again, there are many types of filter which all operate slightly differently. An easy version to apply is the Band pass Butterworth. It is not always necessary to use a band pass filter but it can sometimes be helpful in cleaning up data, particularly where large amounts of gain have been added, typically where the survey took placed over lossy or uneven ground.

Classification Techniques:- Classification is that the process of predicting the category of given data points. Classes are sometimes called as targets/ labels or categories. Classification predictive modeling is that the task of approximating a mapping function (f) from input variables (X) to discrete output variables (y). In machine learning, there are two important categories- Supervised and Unsupervised learning.

Supervised Learning:-When the input variables (x) and an output variable (Y) and an algorithm to seek out the mapping function from the input to the output is supervised learning.

$$Y = f(X)$$

The goal is to approximate the mapping function so well that once you have new input file (x) that you simply can predict the output variables (Y) for that data. It is called supervised learning because the method of algorithm learning from the training dataset is often thought of as an educator supervising the training process.

Unsupervised learning:- Unsupervised Learning is where you simply have input file (X) and no corresponding output variables. The goal for unsupervised learning is to model the underlying structure or distribution within the data so as to find out more about the info. These are called unsupervised learning because unlike supervised learning above there are not any correct answers and there's no teacher. Algorithms are left to their own devises to get and present the interesting structure within the data.

Some of the classification algorithms are:-

ANN:-Artificial Neural Network may be a set of connected input/output units where each connection features a weight related to it started by psychologists and neurobiologists to develop and test computational analogy of neurons. During the training phase, the network learns by adjusting the weights soo on be ready to predict the right class label of the input tuples. There are many network architectures available now like Feed-forward, Convolution, Recurrent etc. The appropriate architecture depends on the application of the model. For most cases feed-forward models give reasonably accurate results and particularly for image processing applications, convolution networks perform better.

SVM:- A support vector machine may be a collection of supervised learning algorithm that use hyper plane graphing to research new, unlabelled data. These machines are mostly employed for classification problems, but also can be used for regression modelling. It also can be used for regression problem. You'll used a SVM when your data has exactly two classes. An SVM classifier data by finding the simplest hyper plane that separates all data point of 1 class from those of the opposite class. The simplest hyper plane for an SVM means the one with the most important margin between the 2 classes. Margin means the maximal width of the slab parallel to the hyper plane that has no interior data point.

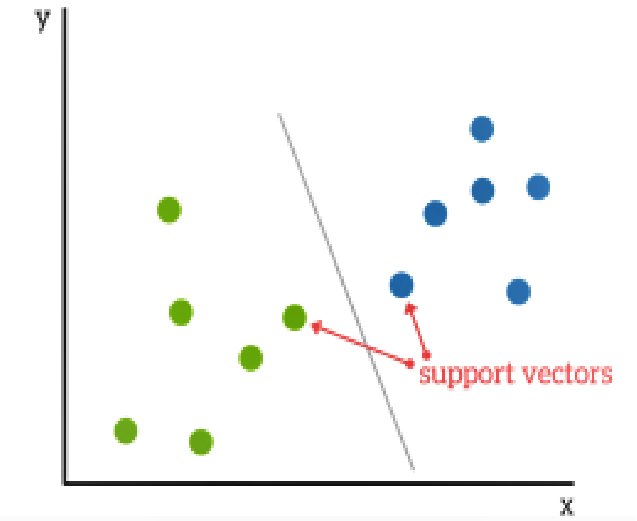

Fig2: Support Vector Machine Graph

Figure 2 shows the co-ordinates of individual observation. Support Vectors are simply the co-ordinates of individual observation. The SVM classifier is a frontier which best segregates the two classes (hyper-plane/ line).

Neuro Fuzzy based classifier:- A classic fuzzy classification rule R_i , which indicate the relation between the classes and the input feature space as follows:

R_i : if x_{s1} is A_{i1} and….. x_{sj} is A_{ij}…..and x_{sn} is A_{in}, then class is C_k

Where x_{sj} represents the j^{th} feature or input variable of the S^{th} sample; A_{ij} represents the fuzzy set of the j^{th} feature in the i^{th} rule; and C_k represents the K^{th} label of class. A_{ij} is identified by the appropriate membership function.

The classifier contains two sub networks i.e. fuzzy self-organizing network and Multi-Layer Perceptron (MLP) in a cascaded way. The feature vector is given as input to fuzzy layer to get pre-classification vector which is given to MLP for classification of test sample. The grouping and clustering of possible nodule pixels in regions based on their membership are done using fuzzy self-organizing layer. This layer would help in differentiating in nodule and non nodule clusters [18].

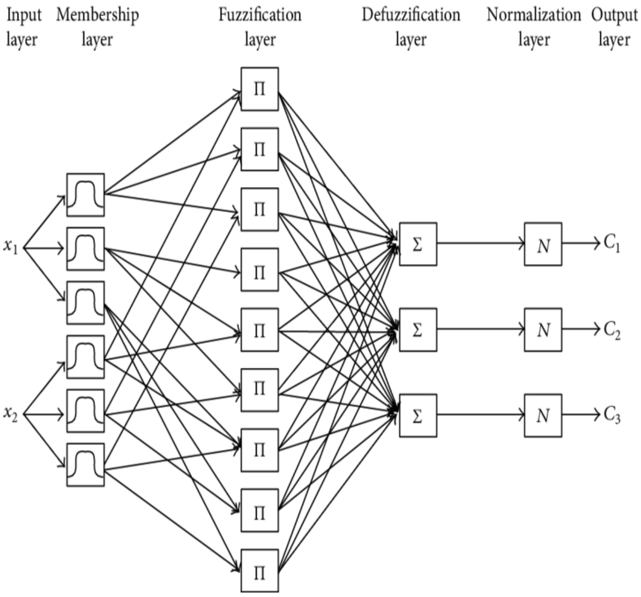

Fig 3: Neuro-fuzzy classifier Architecture

The above figure 3 shows the architecture of neuro-fuzzy classifier with its 6 different layers.

Cellular Automata Filter: - Cellular learning Automata (CLA) model is get from expanding the cellular automata with adding learning automata to each cell. This model is design for arrangements where their components according to experience of themselves and other components experiences are trained, and they have the ability to upgrade their department [19].

IV. Results and Discussion

With the Table 2 we can found a best classifier as well as a feature extraction method which help to classify the lungs images accurately. The above table shows many feature extraction and classification algorithm with its obtaining accuracy and here the maximum accuracy which we get is 98.5% with GLCM feature extraction and ANN classifier.

Table2: Feature extraction and classification algorithms with it are obtain accuracy.

Reference No	Study by	Feature Extraction Method Used	Classification Method Used	Obtain Accuracy
[3]	A. M. Suzan et al.	Scale-invariant feature transform	Bag of features	95%
[14]	K.Gopi et.al	Gray level co-occurrence matrix	Support Vector Machine	92.46%
[15]	S Khobragade et.al	Intensity and discontinuity based method	Artificial Neural Network	92%
[17]	Sheenam Rattan et.al	Gray level co-occurrence matrix	Artificial Neural Network	98.5%
[18]	Anam Tariq et.al	Post Processing Phase	Neuro fuzzy classifier	95%
[19]	Nooshin Hadavi et.al	Band Pass Filter	Cellular Automata	90.9%

V. CONCLUSION

With the above discussion about the features extraction and classification methods of lungs cancer identification, it is conclude that by applying Grey level co-occurrence matrix in feature extraction with classifier as Artificial Neural Network, obtain a highest accuracy. A GLCM feature matrix is generated which can successfully represent a picture with fewer parameters using these properties and Artificial Neural Network is a set of connected input/output units where each connection has a weight associated with it to develop and test computational analogy of neurons. By following the above mention classifier and feature extraction algorithm we can increase the obtain accuracy of the model and can also show the difference between the classifier methods.

REFERENCES

[1] E. Cengil, A. Cinar, "A Deep Learning Based Approach to Lung Cancer Identification", International Conference on Computer Science and Engineering, **2017.**

[2] T.N. Shewaye, and A. A. Mekonnen, "Benign-malignant lung nodule classification with geometric and appearance histogram features" *arXiv preprint arXiv:1605.08350*, **2016.**

[3] A. M. Suzan, and G. Prathibha. "Classification of Benign and Malignant Tumors of Lung Using Bag of Features.",International Journal of Scientific & Engineering Research, **Volume 8, Issue 3, March-2017.**

[4] R.Anirudh, J. J. Thiagarajan, T. Bremer, and H. Kim, "Lung nodule detection using 3D convolutional neural networks trained on weakly labeled data." International Society for Optics and Photonics. 2016 Vol. 9785, p. 978532).

[5] Q.Song, L. Zhao, X. Luo, and X.Dou, "Using Deep Learning for Classification of Lung Nodules on Computed Tomography Images." *Journal of healthcare engineering* Journal of Healthcare Engineering, Volume , Article ID 8314740, **7 pages,2017**

[6] S. M. Salaken, A. Khosravi, A. Khatami, S. Nahavandi, and M.A. Hosen, "Lung cancer classification using deep learned features on low population dataset," In *Electrical and Computer Engineering (CCECE), IEEE 30th Canadian Conference on* (pp. 1-5). IEEE.**2017**.

[7] M.F. Serj, B. Lavi, G. Hoff,. and D. P. Valls, "A Deep Convolutional Neural Network for Lung Cancer Diagnostic," *arXiv preprint arXiv:1804.08170*, **2018**.

[8] E. Cengil, A. Çinar, and Z. Güler. "A GPU-based convolutional neural network approach for image classification." *Artificial Intelligence and Data Processing Symposium (IDAP), 2017 International*. IEEE, **2017**.

[9] D. Jayaraj, S. Sathiamoorthy, "Random Forest based Classification Model for Lung Cancer Prediction on Computer Tomography Images" Second International Conference on Smart Systems and Inventive Technology (ICSSIT 2019).IEEE.

[10] Nidhi S. Nadkarni, Prof. Sangam Borkar," Detection of Lung Cancer in CT Images using Image Processing" Proceedings of the Third International Conference on Trends in Electronics and Informatics (ICOEI 2019) IEEE.

[11] Kuntal Kumar Pal, Sudeep K. S , " Preprocessing for Image Classification by Convolutional Neural Networks" IEEE International Conference On Recent Trends In Electronics Information Communication Technology, **May 2016.**

[12] Michael Blot, Matthieu Cord, Nicolas Thome, "MAX-MIN CONVOLUTIONAL NEURAL NETWORKS FOR IMAGE CLASSIFICATION"IEEE **2016**

[13] Travis Williams, Robert Li, "Advanced Image Classification using Wavelets and Convolutional Neural Networks" 2016 15th IEEE International Conference on Machine Learning and Applications.

[14] K.Gopi, Dr.J.Selvakumar, "Lung tumor Area Recognition and Classification using EK-Mean Clustering and SVM", IEEE **2017**.

[15] Shubhangi Khobragade, Aditya Tiwari, C.Y. Pati1 and Vi kram Narke, "Automatie Deteetion of Major Lung Diseases Using Chest Radiographs and Classifieation by Feed-forward Artifieial Neural Network", 1st IEEE International Conference on Power Electronics. Intelligent Control and Energy Systems (ICPEICES-**2016**).

[16] Sayali Satish Kanitkar, N. D. Thombare, S.S. Lokhande," Detection of Lung Cancer Using Marker-Controlled Watershed Transform", International Conference on Pervasive Computing (ICPC)

[17]Sheenam Rattan,Sumandeep Kaur,Nishu Kansal,Jaspreet Kaur, "An Optimised Lungs Cancer Classification System for Computed Tomography Images", 2017 Fourth International Conference on Image Information Processing (ICIIP).

[18] Anam Tariq , M. Usman Akram and M. Younus Javed*,"*Lung Nodule Detection in CT Images using Neuro Fuzzy Classifier*",* **2013** IEEE.

[19] Nooshin Hadavi, Md.Jan Nordin, Ali Shojaeipour," Lung Cancer Diagnosis Using CT-Scan Images Based on Cellular Learning Automata", **2014** IEEE..

[20] S.K. Vijai Anand," Segmentation coupled Textural Feature Classification for Lung Tumor Prediction", 2010 IEEE.

[21] E. Cengil, A. Çınar , E. Özbay," Image classification with caffe deep learning framework," In Computer Science and Engineering (UBMK), 2017 International Conference on (pp.440-444). IEEE.

[22]P. Mohanaiah, P Sathyanarayana ,L.GuruKumar,"Image Texture Feature Extraction using GLCM Approach", International Journal of Scientific and Research Publications, **Volume 3, Issue 5, May 2013.**

[23] Rahul Meena, Vighnesh Menon, Vivek Solavande, "Lung Image Classification Using Convolutional Neural Network And Prediction of Different Diseases", International Journal of Computer Sciences and Engineering, **Vol.8, Issue.4, pp.2347-2693, 2020.**

[24] Zarli Cho1 , Khin Myo Kyi , Kyi Thar Oo, "Image Classification based on Feature Extraction with AlexNet Architecture"International Journal of Computer Sciences and Engineering, **Vol.8, Issue.4, pp. 2347-2693, 2020.**