A Brief Survey on Lung Cancer Detection Using Convolutional Neural Network

Mahesh N ,Dr.Keshava Prasanna

Research Scholar,Department of CSE,CIT,Gubbi,Tumkur,Karnataka,India.
Research Supervisor,Department of CSE,CIT,Gubbi Tumkur,Karnataka,India.

mahe.sittumkur@gmail.com, keshavaprasanna2013@gmail.com

Article History: **Received:** 19.05.2023 **Revised:** 20.06.2023 **Accepted:** 10.07.2023

Abstract

Lung most cancers is the main purpose for cancer-related death. Lung most cancers can provoke in the windpipe, major airway or lungs. It is brought about by way of unchecked increase and unfolds of some cells from the lungs. People with lung disorder such as emphysema and preceding chest troubles have greater danger to be identified with lung cancer. Over utilization of tobacco, cigarettes and beedis, are the fundamental chance thing that leads to lung most cancers in Indian men; however, amongst Indian women, smoking is now not so common, which point out that there are different elements which lead to lung cancer. Other danger elements consist of publicity to radon gas, air-pollutions and chemical compounds in the workplace.

Lung most cancers detection at early stage has come to be very vital and additionally very handy with photograph processing and deep getting to know techniques. In this find out about lung affected person Computer Tomography (CT) scan pix are used to discover and classify the lung nodules and to notice the malignancy stage of that nodules. In this challenge mainly the usage of CNN algorithm to become aware of Lung most cancers from CT-SCAN pics and to educate CNN have CT-SCAN photographs dataset.

Keywords: Convolutional Neural Network (CNN), Machine Learning, Lung Cancer, Histopathological Image

1. INTRODUCTION

It is most common in smokers accounting 85% of cases among all. So many Computer Aided Diagnosis (CAD) Systems are developed in recent years. Detection of lung cancer at early stage is necessary to prevent deaths and to increase survival rate. Lung nodules are the small masses of tissues which can be cancerous or noncancerous also called as malignant or benign. Benign tissues are most commonly non-cancerous and does not have much growth where malignant tissues grows very fast and can affect to the other body parts and are dangerous to health.

Various tests like imaging sets (x-ray, CT scan), Sputum cytology, and tissue sampling (biopsy) are carried out to look for cancerous cells and rule out other possible conditions. While performing the biopsy, evaluation of the microscopic histopathology slides by experienced pathologists is indispensable to establishing the diagnosis [3], [4], [5], and defines the types and subtypes of lung cancers [6]. For pathologists and other medical professionals diagnosing lung cancer and the types is a time-consuming process. There is a significant change the cancer types are misdiagnosed, which directs to incorrect treatment and may cost patients' lives.

Early prediction of lung cancer should play a pivotal role in the diagnosis process and for an effective preventive strategy. There are many techniques to diagnosis lung cancer, such as Chest Radiograph (x-ray), Computed Tomography (CT), Magnetic Resonance Imaging (MRI scan) and Sputum Cytology. However, most of these techniques are expensive and time consuming. Most of these techniques are detecting the lung cancer in its advanced stages, where the patient's chance of survival is very low. Therefore, there is a great need for a new technology to diagnose the lung cancer in its early stages. Image processing techniques provide a good quality tool for improving the manual analysis.

Using the significant pattern prediction tools for a lung cancer prediction system will develop. This lung cancer risk prediction system should prove helpful in detection of a person's predisposition for lung cancer. Therefore early prediction of lung cancer should play a pivotal role in the diagnosis process and for an effective preventive strategy.

There are many techniques to diagnose lung cancer such as CT-SCAN, MRI, X-RAY. These techniques are expensive and time consuming. Most of these techniques are detecting the lung cancer in its advanced stages, where the patients' chance of survival is very low. Therefore, there is a great need for a new technology to diagnose the lung cancer in its early stages. The study on lung cancer can be used to detect lung cancer in early stages. The given study for early detection and prediction of lung cancer system which is easy cost effective and time saving.

This study for lung cancer detection and prediction system helps to detect the lung cancer in its early stage and also to predict the lung cancer. Therefore the survival rate of patient will increase. The purpose behind to designing this system is to predict and detect the lung cancer in its early stage on the basis on some factors and thresholding. The decrease and the number of rule for testing in this system is to be considered. This system reduce the time and cost required for various excessive medical test. The propose system is on web based due to this a rural site patient directly communicate with doctors and doctors will try to solve their questions (problems).

Machine Learning (ML) is a subfield of Artificial Intelligence (AI) that allows machines to learn without explicit programming by exposing them to sets of data allowing them to learn a specific task through experience [7][8]. In previous research papers, most of the authors considered using x-rays, CT scans images with machine learning techniques such as Support Vector Machine (SVM), Random Forest (RF), Bayesian Networks (BN), and Convolutional Neural Network (CNN) for lung cancer detection and recognition purpose. Some papers also considered using histopathological images, but they distinguish between carcinomas and non- carcinomas images and with lower accuracy. This research paper has considered using Convolutional Neural Network (CNN) architecture to classify the benign, Adenocarcinoma, and squamous cell carcinomas. We have not found other papers using the CNN model to classify only the given three different histopathological images and the given model's accuracy.

It has been established that malignant tumors as well as metastases, of almost all histological types, can regress spontaneously although certain histological types regress more frequently than others. Various causes thereof include apoptosis, the immune system and particular conditions of the tumor microenvironment.

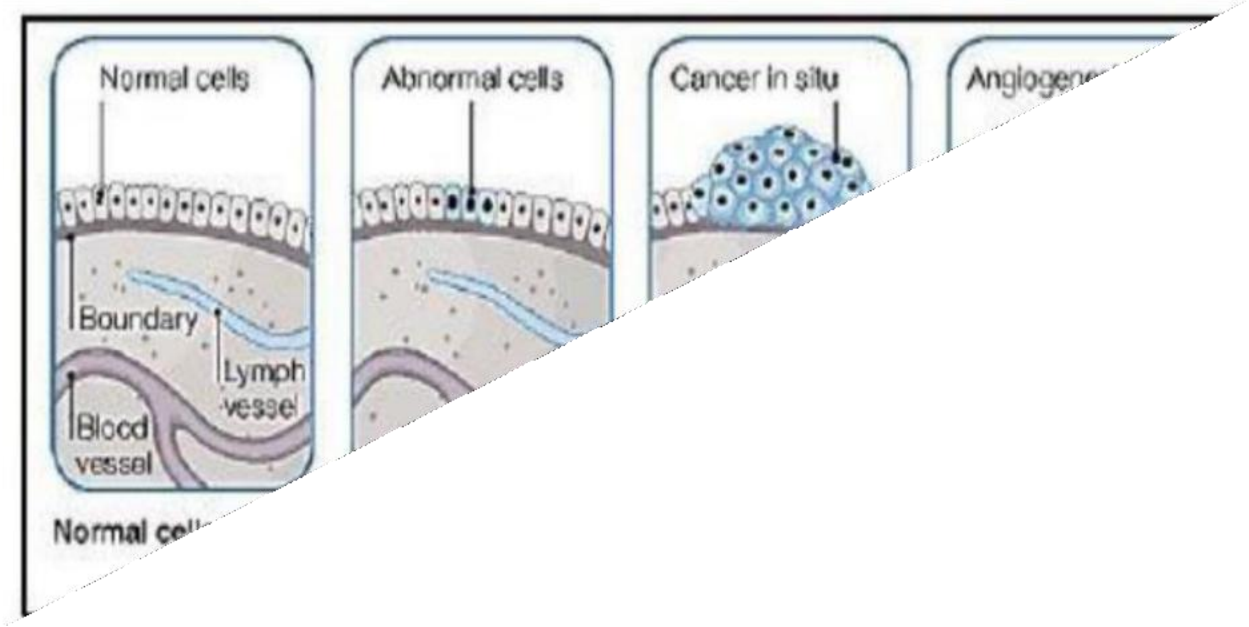

Figure 1: The beginning of cancer- N Bhaskar[1]

SCOPE:

For medical imaging so many different types of images are used but computer Tomography (CT) scans are generally preferred because of less noise. Deep learning is proven to be the best method for medical imaging, feature extraction and classification of objects. Several types of deep learning architectures are introduced by so many researchers to classify the lung cancer.

1.1 Contribution of the Research Work

1. A novel hybrid deep learning based model is proposed for theearly detection of lung cancer using CT scan images. Theproposed architecture has been trained with LIDC datasetsand performance metrics have been calculated and comparedwith other existing models.
2. The proposed architecture introduces the capsule network'sbetter segmentation and transfer learning for featureextraction. Also, the proposed fusion algorithm canincrease the high diagnosis rate.

3. The whale optimization algorithm is proposed for trainingthe features obtained from the hybrid fusion of saliency mapsand capsule networks. The feed-forward layers are designedbased on the principle of Extreme Learning Machines (ELM).The rest of the paper is organized as follows: Section-IIpresents the related works proposed by more than one author.The working mechanism of the saliency maps, CNN layers, antlion optimization, and feedforward networks are presented inSection-III. The dataset descriptions, experimentations, results,findings, and analysis are presented in Section-IV. Finally, thepaper is concluded in Section-V with future enhancements.

2. LITERATURE REVIEW

Cancer causes about one in six deaths every year [1][2] and lung cancer stands at the top of all of this as it is responsible for 1.76 million deaths up to 2016. [1] Early detection of cancer can provide a suitable treatment to not just prolong life but also save a patient’s life and hence increase the survival rate. [1][2][3][4]

The journal paper [1] published by Muthazhagan B, Ravi T, Rajinigirinath D (2021) states that with the aid of current lung cancer prediction technologies, predicting and detecting lung cancer at an early stage is a difficult challenge. An early lung tumor prediction might extend a person's life by one to five years. They created a Support Vector Machine based classification model which provided about 98% prediction accuracy in a small amount of time. However, the images were merely classified into ‘abnormal’ or ‘normal’ and did not take into account the various stages [Stage 0 – stage IV] which is what this project aims to improve on. The paper [2] proposed by Masud M, Sikder N, et al. (2021) uses a CNN based model for classifying the image into one of five kinds: colon adenocarcinoma, benign colonic tissues, lung adenocarcinomas, lung squamous cell carcinomas and benign lung tissues.

While a peak accuracy of 96.33% has been achieved in the classification, the authors state that two out of five classes can have much improved performance with further experimentation. The dataset used is Histopathological and Histopathology is the microscopic examination of a biopsy which is an invasive process. Our approach prefers to work on CT scans which is a non-invasive mechanism to detect cancer.

Sajja T, Devarapalli R, et al. (2019) [3] published a paper which worked on detecting lung cancer using the pre-trained CNN model called Google-Net. The deployed 60% of all neurons in the drop out layers to prevent overfitting and achieved a simplified and sparse network for classifying the CT images into benign or malignant. The model still requires testing on various dropout ratios to check for better performance accuracy. Our approach aims to construct a simplified CNN model to classify cancer along with providing medical information costs.
Tripathi P, Tyagi S, et al. (2019) [4] published a paper in which they attempt to detect lung cancer using four different segmentation techniques of image processing.

They conclude that marker-controlled watershed segmentation provides the most accurate results. Through the comparative analysis, it is found that CT scans tend to provide the best chance at detecting cancer and should be the preferred means to do the same. Hence, to use Deep Learning on CT scans to classify the various stages.

Nasrullah Nasrullah et al. (2019) [5] study focuses on developing a model that can detect cancerous nodules using CT images. They opt to employ 3D CNN after some research because of its proven performance in image analysis. To further identify the condition as benign or malignant, they use 3D MixNet to extract nodule features, which are then classified using Gradient Boosting Machine (GBM). The proposed model was validated using the free response receiver operating characteristic (FROC) evaluation matrix to obtain a FROC score of 94.21%. The suggested model outperformed all other models in terms of computational cost and desired output accuracy.

Siddharth Bhatia et al. (2019) [6] present a method for detecting lung cancer using deep residual learning. They offer a series of preprocessing strategies for extracting cancer-vulnerable lung features using UNet and ResNet models. They examine the likelihood of predicting carcinogenic CT scans by comparing the effectiveness of classifiers such as Random forest and XGBoost.

When the authors combine the two classifiers, they get the greatest accuracy of 84%. The constraint in this case is that the best achievable accuracy may have been higher.

SurenMakaju et al. (2018) [7] made a comparison of many probable cancer detection approaches and ranked them in order of effectiveness. They decide to upgrade that model to achieve even higher accuracy by selecting the current best approach from their survey of articles.

The Median and Gaussian filters were used in the pre-processing stage, and the data was then segmented using the Watershed algorithm. They went on to use support vector machines to identify diagnosed cancerous nodules as benign or malignant. This upgraded model outperformed the previous best model by 5.4%, with an accuracy rate of 92 %. The model's sole flaw is that it does not differentiate between cancer stages (I to IV).

Inspired by the AlphaGo system, Ali I et al. (2018) [8] developed a deep learning algorithm that takes a CT image and perceives it as a collection of states, producing a classification of whether or not a malignant nodule is present. They employ a Reinforcement Learning algorithm that improves with time and with more data. Their research shows that the model's training data has a high accuracy of 99.1%, however the validation data has a low accuracy of 64.4 %. The model appears to be over fitted as a result of this. The authors suggest that because this is the only flaw, the constraint can be solved with more data.

The authors W. Ausawalaithong, A. Thirach, S. Marukatat, and T. Wilaiprasitporn [9] used deep learning with a transfer learning approach to predict lung cancer from the chest X-ray images obtained from different data sources. Image size of 224X224 with 121-layer Densely Connected Convolutional Network (DenseNet-121) and a single sigmoid node was applied in a fully connected layer. The proposed model achieved 74.43±6.01% mean accuracy, 74.96±9.85% of mean specificity, and 74.68±15.33% mean sensitivity for different image source dataset.

T. Atsushi, T. Tetsuya, K. Yuka, and F. Hiroshi [10] applied Deep Convolutional Neural Network (DCNN) on cytological images to automate lung cancer type classification. They considered Small cell carcinoma, Squamous cell carcinoma, Adenocarcinoma images in their dataset. The DCNN architecture of 3 convolution and pooling layers and 2 fully connected layers with dropout 0f 0.5 were used. The model developed was able to achieve the overall accuracy of 71.1%, which is quite low.

W. Rahane, H. Dalvi, Y. Magar, A. Kalane, and S. Jondhale [11] proposed using image processing and machine learning (Support Vector Machine) for lung cancer detection on computed tomography (CT) images. Image processing like grayscale conversion, noise reduction, and binarization was carried out. Features like area, perimeter, and eccentricity from the segmented image region of interest were fed to the support vector machine (SVM) model.

M. Šarić, M. Russo, M. Stella, and M. Sikora [12] proposed CNN architectures implementing VGG and ResNet for lung cancer detection using whole side histopathology images, and the output was compared using the receiver operating characteristic (ROC) plot. Patch level accuracy of 0.7541 and 0.7205 was obtained for VGG16 and ResNet50 respectively which is quite low. The authors explained that the given models' low accuracy was due to large pattern diversity through different slides.

The authors S. Sasikala, M. Bharathi, B. R. Sowmiya [13], proposed using CNN on CT scan images to detect and classify lung cancer. They used MATLAB for their work and has two phases in training to extract valuable volumetric features from input data as the first phase and classification as the second phase. Their proposed system could classify the cancerous and non-cancerous cells with 96% accuracy.

SRS Chakravarthy, R. Harikumar [14], used Co-Occurrence Matrix (GLCM) and chaotic crow search algorithm (CCSA) for feature selection on computed tomography (CT) and applied probabilistic neural network (PNN) of the classification task. They found that the PNN model build on CCSA features performed better with 90% accuracy.

In De Bruijne (18), the presented framework looked at the mostup-to-date lung cancer detection and diagnosis methods. Usingstandardized databases LIDC-IDRI, LUNA 16, and Super Bowl Dataset 2016, the newest lesion detection, identification, anddetectors are acquainted with labeled models. According to theauthor Jindal et al. (19), these are the most common and typical threshold CT data considered for diagnosing. The authors inNalepa and Kawulok (20) developed the modified-CNN in orderto recognize the tumor cells in the lung regions with the segmented images. The ACM method has been used forsegregating the tumor region initially and identifying cancer ornormal cells.

The label-free techniques do not injure cells or cause effects oncell structure or intrinsic features. To enhance cell identificationusing recorded optical profiles, this study combined advancementsin optical coherence tomography with Pronymethods. In Ganesan et al. (21), the framework finds signature genes by improvingTobacco Exposure Pattern (TEP) Prediction model and revealingtheir interaction connections at many biological levels. TTZ.

Kasinathan et al. (22) is a new way to extract core features anduse them as an input variable in the TEP classification model.With two distinct LUAD datasets used to train and evaluate the TEP classification model, 34 genes were identified as nicotineassociatedmutation signature genes, with an accuracy of 94.65%for training data and 91.85% for validation data.

The researcher examined tissue samples and devised acategorization method to discriminate between five types ofpulmonary and colorectal tissues (two benign and threemalignant). According to the observations, the suggestedapproach can detect tumor cells up to 96.33% of the time (23).

The framework presented in Suzuki (24) described how to usecomputer-assisted diagnostics to assess EGFR mutation status,including gathering, evaluating, and merging multi-type interdependence characteristics. This research uses a new hybridnetwork model based on CNN-RNN architecture. CNN is used toextract image quantitative properties, and the link between different types of features is modeled. Their study indicate thatmulti-type dependency-based feature representations beat single typefeature representations (accuracy = 75%, AUC = 0.78) whencompared to conventional features extracted.

The 3D_Alex Net unsupervised learning model (25) wasintroduced for lung cancer detection. The 3D CNN is a highlypredictive architecture with an improved steepest descent input signal that increases the appearance of tumor tissues. The LUNAdatabase is used to assess the proposed Alex Net detectiontechnique to an existing 2D CNN training classifier. Due to a lack of testing data, the proposed model is unsuccessful, with just10% of the training database being utilized.Tajbakhsh and Suzuki (26) examined the performance ofCNNs and MTANNs for detecting and classifying lung nodules.Achieving 100% sensitivity and 2.7 false positives per patient,MTANN exceeds the top performing CNN (AlexNet) in their testing. The MTANNs achieved an accuracy of 0.88 in classifyingnodules as benign or malignant.Gu et al. (27) suggested a unique 3D-CNN CAD system for lungnodule detection. They used a multiscale technique to improve thesystem's detection of nodules of varying sizes. The suggested CAD system considers preprocessing, which is common in standaloneCAD systems. It uses volume segmentation to create ROI cubes for3D-CNN classification. After categorization, DBSCAN was used toblend adjacent regions that could be from the same nodule. Largerscale cubes have lesser sensitivity (88%) but an average of one falsepositive per patient, according to the LUNA16 dataset.

The multi-section CNN model suggested by Sahu et al. (28)uses multiple view sampling to classify nodules and estimatemalignancy. Their proposed model is faster than the widely utilized 3D-CNNs. To develop their system, they employed pretrained Mobile Net networks and sample slices extracted in variousdirections. On the LUNA2016 dataset, the suggested model had asensitivity of 96% and an AUC of 98%. They estimate the classlikelihood of malignancy using a logistic regression model. Itestimated malignancy with 93.79% accuracy. Because it is solight, it can be used on smaller devices like phones and tablets.

Deep3DSCan was proposed by Bansal et al. (29). To do so, theyapplied a deep 3D segmentation technique on CTs. The ResNet based model was trained using a combination of deep fine-tuned residual network and morphological features. The LUNA16 datasetwas utilized for training and testing. The proposed architectureachieved an F1 score of 0.88 in segmentation and classification tasks.

In Jothi et al. (30), the framework designed a controlled CNNclassifier for patients with lung cancer to detect potential adenocarcinoma (ADC) and squamous cell carcinoma (SCC).CNN has already been verified using authentic Non-SCLC patientinformation from preliminary phase afflicted subjects collected at Ramana et al. Lung Cancer Prediction Using DL Frontiers in Massachusetts General Hospital (31). In the record, there are 311data phases that have been collected. The created CNN, which is aVGG system training predictor, only had a 71% AUC predictiveperformance, which was insufficient. The VGG CNN model's flaw is that it hasn't been preprocessed for background subtraction orimage reconstruction fragmentation, which increases the predictiveaccuracy.

In Kasinathan and Jayakumar (32), the new cloud-basedtumor recognition model was developed. The author analyzedvarious standard dataset "CT-scans and PET-scans" forsegmenting the ROC and for recognizing the tumor. In Jakimovski and Davcev (33), the framework proposes a noveldeep learning method based on binary particle swarmoptimization with a decision tree (BPSO-DT) and CNN toidentify the malignant or normal cells in the lung region usingthe genetic features (34).

All researchers have aim to develop such a system which predict and detect the cancer in its early stages. Also tried to improve the accuracy of the Early Prediction and Detection system by preprocessing, segmentation feature extraction and classification techniques of extracted database. The major contributions of the research are summarized below.

T. Sowmiya, M. Gopi, M. New Begin, L.Thomas Robinson - In this paper they described Cancer as the most dangerous diseases in the world. Lung cancer is one of the most dangerous cancer types in the world. These diseases can spread worldwide by uncontrolled cell growth in the tissues of the lung. Early detection of the cancer can save the life and survivability of the patients who affected by this diseases. In this survey paper several aspects of data mining procedures which are used for lung cancer prediction for the patients. Data mining concepts is useful in lung cancer classification. Major part also reviewed the aspects of ant colony optimization (ACO) technique in data mining. Ant colony optimization helps in increasing or decreasing the disease prediction value of the diseases.

Ada[1], Rajneet Kaur - In this paper uses a computational procedure that sort the images into groups according to their similarities. In this paper Histogram Equalization is used for preprocessing of the images and feature extraction process and neural network classifier to check the state of a patient in its early stage whether it is normal or abnormal. After that predict the survival rate of a patient by extracted features. In this paper Neural Network Algorithm is implemented using open source and its performance is compared to other classification algorithms.

DasuVaman Ravi Prasad - In this paper image quality and accuracy is the core factors of this research, image quality assessment as well as improvement are depending on the enhancement stage where low pre- processing techniques is used based on Gabor filter within Gaussian rules. Following the segmentation principles, an enhanced region of the object of interest that is used as a basic foundation of feature extraction is obtained. Relying on general features, a normality comparison is made. In this research, the main detected features for accurate images comparison are pixels percentage and mask labeling.

S Vishukumar K. Patela and PavanShrivastavab - In this paper authors mostly focus on significant improvement in contrast of masses along with the suppression of background tissues is obtained by tuning the parameters of the proposed transformation function in the specified range. The manual analysis of the sputum samples is time consuming, inaccurate and requires intensive trained person to avoid diagnostic errors. The segmentation results will be used as a base for a Computer Aided Diagnosis (CAD) system for early detection of cancer, which improves the chances of survival for the patient. In this paper, authors proposed gabor filter for enhancement of medical images. It is a very good enhancement tool for medical images.

Fatma Taher1,*, Naoufel Werghi1, HussainAl-Ahmad1, Rachid Sammouda2 (2012) - This paper presents two segmentation methods, Hopfield Neural Network (HNN) and a Fuzzy C-Mean (FCM) clustering algorithm, for segmenting sputum color images to detect the lung cancer in its early stages. The manual analysis of the sputum samples is time consuming, inaccurate and requires intensive trained person to avoid diagnostic errors. The segmentation results will be used as a base for a Computer Aided Diagnosis (CAD) system for early detection of lung cancer which will improve the chances of survival for the patient. However, the extreme variation in the gray level and the relative contrast among the images make the segmentation results less accurate, thus applied a thresholding technique as a pre-processing step in all images to extract the nuclei and cytoplasm regions, because most of the quantitative procedures are based on the nuclear feature. In this study, we used 1000 sputum color images to test both methods, and HNN has shown a better classification result than FCM, the HNN succeeded in extracting the nuclei and cytoplasm regions.

SL.NO	Author Name and Year	Title of Paper	Methodology	Limitations/ Conclusions
1	Muthazhagan B, Ravi T, Rajinigrinath D - 2021	An Enhanced Computer assisted Lung Cancer-Based Image Retrieval and Data Mining Techniques. [1]	Support Vector Machine image classification algorithm	The malignancy is classified as ‘Normal’ and ‘Abnormal’ not as Stages 1-4
2	Masud M, Sikder N, et al. – 2021	A Machine Learning Approach to Diagnosing Lung and Colon Cancer Using a Deep Learning Based Classification Framework [2]	Digital Image Processing techniques with CNN	Dataset uses microscopic cells images rather than CT/MRI scans
3	Sajja T, Devarapalli R, Kalluri H- 2019	Lung Cancer Detection Based on CT Scan Images by Using Deep Transer Learning [3]	A deep neural network based on Google – Net.	Over fitted data causing the need for max dropout ratio.
4	TRIPATHI p, Tyagi S, Nath M	Comparative Analysis of	Comparative Analysis	Marker controlled watershed segementation

	- 2019	Segmentation Techniques for Lung Cancer Detection[4]	image Segmentation techniques	provides more accurate results.
5	Nasrullah N, Sang J, Alam MS, Mateen M, Cai B, HUu H-2019	Automated Lung Nodule Detection and Classification Using Deep Learning Combined with Multiple Strategies[5]	Two 3D CNN with CMixNet architectures	3D CMixNet had better accuracy feature exploitation than other models compared with.
6	Bhatia S, Sinha Y,GOEL L- 2019	Lung Cancer Detection: A Deep Learning Approach [6]	Deep residual networks with XGBoost and Random Forest classifers and ensemble	The highest accuracy was84% Using an ensemble of both models tried which still a comparatively low accuracy.
7	Makaju S, Prasad PW, et al -2018	Lung Cancer Detection using CT Scan Images [7]	Watershed algorithm with SVM.	Classification of different stages of cancer is not done.
8	Ali I, Hart GR, et al. -2018	Lung Nodule Detection via Deep Reinforcement Learning [8]	Reinforcement Learning algorithm	The model is overfit as training accuracy obtained was 99.1% whereas the testing accuracy was 64.4%.

Objectives of this system are as follows:

- To decrease the number of rules for testing.
- To reduce the time and cost required for various excessive Medical Tests.

- To increase the accuracy of performance of Lung Cancer Prediction and Detection System.
- Use less number of attributes for prediction of Cancer.
- Early stage detection of cancer.
- Increasing the survivability of the patient more than 5 years.

III. METHODOLOGY TO BE FOLLOW

Diagnosis of lung cancer includes the following stages:

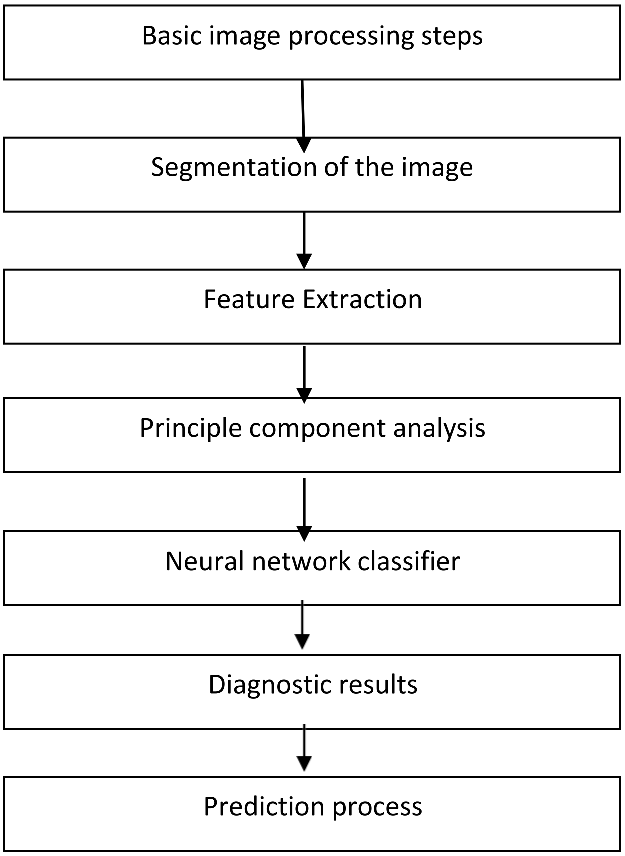
Basic image processing steps
Segmentation of the image
Feature Extraction
Principle component analysis
Neural network classifier
Diagnostic results
Prediction process

Figure 2: flow chart of the proposed method

1. Images acquisition:

Primarily, cancer and non-cancer patient's data or CT-Scan images will be collected from different diagnostic centers. The digitized images are stored in the DIACOM format with a resolution of 8 bits per plane [1].

2. Preprocessing:

The image Pre-processing stage in this system begins with image enhancement which aims to improve the interpretability or sensitivity of information included in them to provide better input for other programmed image processing techniques.

Image enhancement techniques can be divided into two wide types: Spatial domain methods and frequency domain methods. On the other hand, when image enhancement techniques are used as preprocessing tools for other image processing techniques, the quantifiable measures can determine which techniques are most suitable. In the image enhancement stages will be using the Histogram Equalization.

The pre-processing of image aims for selective elimination of the redundancy in scanned images without affecting the original image, this play a vital role in the diagnosis of lung cancer. Therefore, Histogram- Equalization becomes the crucial step in preprocessing. Hence, each image is preprocessed to enhance its superiority.

3. Converting in to grey scale:

The average pixel values (ranging from 0-255) of the primary colours which are red green and blue (popularly referred to as RGB) are combined. The luminous intensity of each colour band (which is 24 bits) is combined into a reasonable approximated grayscale value(8 bits)

4. Segmentation:

Image segmentation is a crucial process for most image analysis consequent tasks. Especially, most of the existing techniques for image description and recognition are highly depend on the segmentation results. Segmentation splits the image into its constituent regions or objects. Segmentation of medical images in 2D has many beneficial applications for the medical professional such as: visualization and volume estimation of objects of concern, detection of oddities, tissue quantification and organization and many more.

The main objective of segmentation is to simplify and change the representation of the image into something that is more significant and easier to examine. Image segmentation is usually used to trace objects and borders such as lines, curves, etc. in images. More accurately, image segmentation is the process of allocating a label to every pixel in an image such that pixels with the same label share certain pictorial features.

The outcome of image segmentation is a set of segments that collectively cover the entire image, or a set of edges extracted from the image i.e. edge detection. In a given region all pixels are similar relating to some distinctive or computed property, such as texture, intensity or color. With respect to the same characteristics adjacent regions are significantly different.

One of two basic properties of intensity values Segmentation algorithms are based on: discontinuity and similarity. In the first group partition the image based on abrupt changes in intensity, such as edges in an image. The next group is based on segregating the image into regions that are alike according to a predefined criterion. Histogram thresholding methodology comes under this group.

5. Feature Extraction:

Image features Extraction stage is a crucial stage that uses algorithms and methods to detect and separate various preferred portions or shapes of an inputted image.

The following two methods are used to predict the probability of lung cancer presence: binarization and GLCM, both methods are based on facts that strongly related to lung anatomy and information of lung CT imaging.

A. Binarization Approach

For detection of cancer binarization approach has been applied for detection of cancer. In binarization to extract the number of white pixels and check them against some threshold to check the normal and abnormal lung cells. After this process the condition is check whether number of white pixels of a new image is less than the threshold then it indicates that the image is normal, or else if the amount of the white pixels is greater than the threshold, it specifies that the image in abnormal. Merging Binarization and GLCM methods together will lead us to take a decision whether the case is normal or abnormal.

B. GLCM (Grey Level Co-occurrence Method)

The GLCM is a process of tabulating different combinations of pixel brightness values called as grey levels which occurs in an image. In this first step is to create gray-level co-occurrence matrix from image in MATLAB.

C. Masking Approach

Inside lungs masses are appeared as white connected areas inside ROI (lungs), masking approach depends on this. As they increase the percent of cancer presence increases. Also combining Binarization and Masking approaches together will help us to take a decision on whether the case is normal or abnormal according to the mentioned assumptions in the previous two approaches, it can make a conclusion that if image has number of black pixels greater than white pixels then that image is normal or otherwise it can say that the image is abnormal.

5. PCA (Principle Component Analysis)

PCA is a technique to normalize the data in image. Real-world data sets generally display associations among their variables. These associations are frequently linear, or at least practically so, making them agreeable to common examination techniques. One such technique is principal component analysis ("PCA"), which rotates the original data to new coordinates, making the data as "even" as possible. The features mined are delivered to the PCA data mining for better sorting [1].

The following steps takes place in PCA:-

i. Calculating the mean and standard deviation of the features in the image.

ii. Subtracting the sample mean from each observation, and then dividing by the sample standard deviation. This scales and centers the data.

iii. Then to calculate the coefficients of the principal components and their relevant changes are done by finding the Eigen function of the sample covariance matrix.

iv. This matrix holds the coefficients for the principal constituents. The diagonal elements store the modification of the relevant principal constituents. It can mine the diagonal.

v. The maximum variance in data results in maximum information content which is required for better classification [1].

6. Neural Network Classifier

Supervised feed-forward back-propagation neural network ensemble used as a classifier tool. Neural network contrasts in different means from traditional classifiers like Bayesian and k – nearest neighbor classifiers. Linearity of data is one of the major variances. Other existing classifiers like Bayesian and k – nearest neighbor entails linear data to work properly. But neural network works as well for nonlinear data because it is simulated on the reflection of biological neurons and network of neurons.

Training the neural network with wide range of input data will increase the detection accuracy, in other words the system will get biased with a small set of data or large set of similar data. Hence neural network classifier needs a large set of data for training and also it is time consuming to train to reach the stable state. But once it is trained it works as fast and quick as biological neural network by transmitting signals as fast as electrical signals.

Input layer, internal hidden layer and output layer are the three layers of the architecture of the neural network. The nodes in the input layer are linked with a number of nodes in the internal hidden layer. Each input node connected to each node in the internal hidden layer. The nodes in the internal hidden layer may connect to nodes in another internal hidden layer, or to an output layer. And the output layer consists of one or more response variables.

Following are the general Steps performed in Neural Network Classifier:-

- Creating feed-forward back propagation network.
- Training neural network with the already available samples and the group defined for it.

- The input image mined PCA consistent data as the test samples, fires the neural network to check whether the particular selected input sample has cancer or not.
- From the results which are obtained from the neural network and the samples trained in network classification rate is calculated using some mathematical formulas.

7. Diagnosis Result

After completion of all the processes in the last stage i.e. in the diagnosis stage or in diagnosis result the proposed system show whether the image is in normal or in abnormal state.

8. Prediction process

There is no remedy for cancer after completely affected. Death is inevitable. So the ability to predict Lung cancer plays an important role in the diagnosis process. In this paper the proposed an effective Lung cancer prediction system based on data mining. This lung cancer risk prediction system should prove helpful in detection of a person’s predisposition for lung cancer. it will be considering various risk factors which includes-age, gender, hereditary, previous health examination, use of ant hypersensitive drugs, smoking, food habit, physical activity, obesity, tobacco, genetic Risk, environment, mental trauma, uptake of red meat, balance diet, hypertension, heart disease, excessive alcohol, radiation therapy and chronic lung diseases. Various algorithms will be used such as decision tree algorithm for the prediction process.

V. RESULTS AND DISCUSSIONS

In order to prepare the image for segmentation, pre-processing of the image was done by contrast enhancement and median filtering. Median filter was used for noise removal. Contrast enhancement was performed. The number of features reduced by feature selection methods with Genetic algorithm based Initialization for the optimization of results. Reducing the number of features of dataset is important .All methods were successful in reducing the number of features. The image is resized to 128 by 128 and the average, min and max of 16x16 windows is used to extract for each angle. Thus generating 64 features of min, max and avg. Then the average with respect to each angle is taken thus generating 192 features. The Fuzzy c means clustering is done to cluster the images and segmentation of the images are done. CFS and IG is used to rank this feature and classification accuracy is used to measure with 40, 60, 80, 100 features using genetic algorithm. The classification accuracy of datasets with 10-fold cross validation for finding the accuracy of the images are computed.

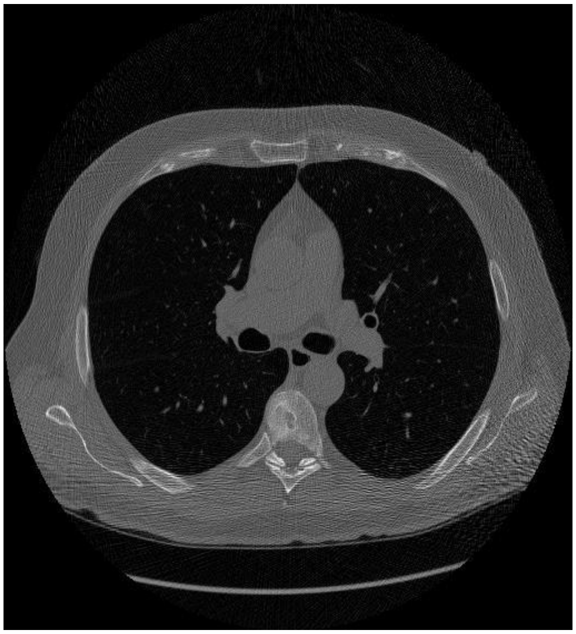

Figure 3: input of the lung cancer image

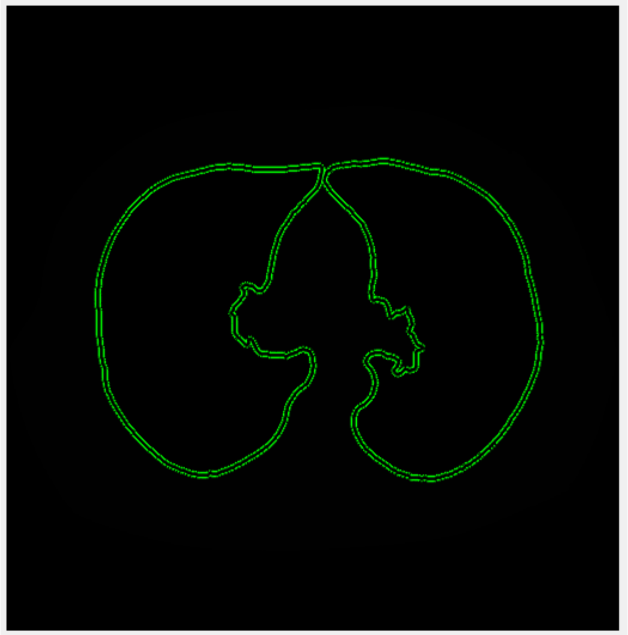

Figure 4: detection of the cancer image

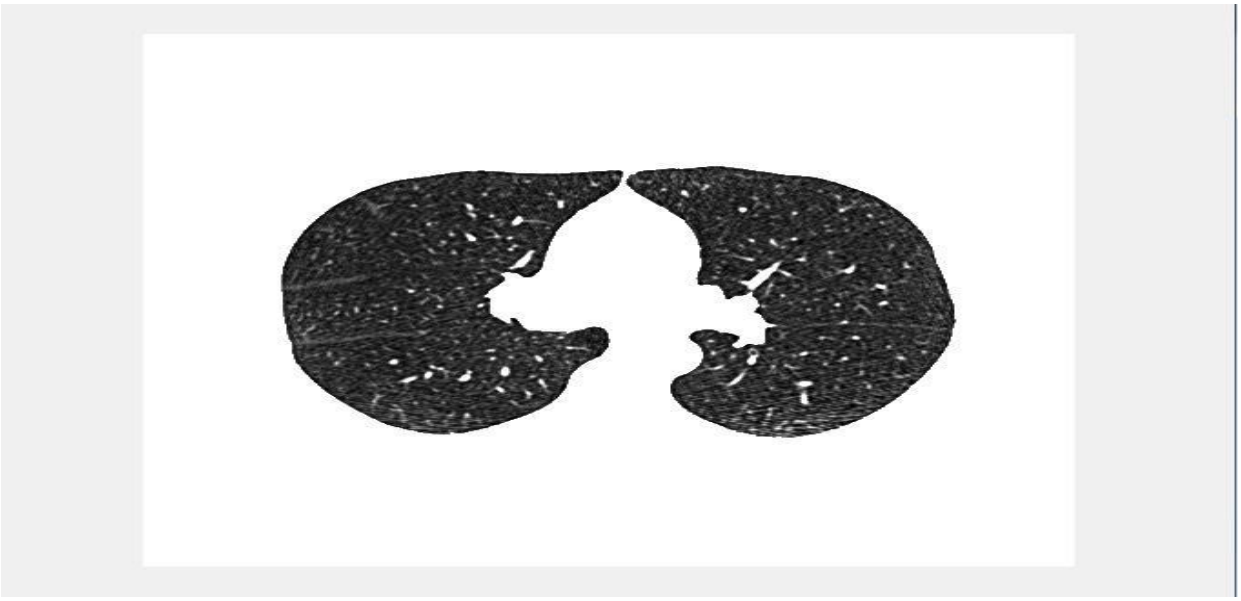

Figure 5: output detected cancer image

VI. CONCLUSION

A technique for quick detection of lung cancer using CT scans was proposed in this research. We discovered that the random forest classifier and regressor gave more accurate results than other algorithms when it examined different machine learning techniques from the survey. In addition, Convolutional Neural Networks with class weights provide reliable results for categorizing CT scan pictures by overcoming unbalanced data issues. When compared to previous transfer learning models, the GoogLeNet model successfully predicts the malignant tumor level in less time. As a result, the proposed model overcomes the disadvantages and provides an application in which the user may obtain information ranging from a basic level, such as symptom prediction. Then it allows lung cancer stage classification using CT scans, and presents insurance costs. The user may obtain all of the information they want from a single application in less time.

Using the combination of neural network classifier along with binarization and GLCM will increase the accuracy of lung cancer detection process. This system will also decrease the cost and time required for cancer detection. Also if the patient is not detected with the lung cancer the system will proceed further for the prediction process. As this system will be available online, patients from remote areas can also avail its benefits. So this system is beneficial for huge number of people all over the world. Also tests required for cancer detection is required.

REFFERENCES

1. G. A. Silvestri, et al. "Noninvasive staging of non-small cell lung cancer: ACCP evidence-based clinical practice guidelines (2nd edition)." Chest vol. 132, 3 Suppl (2007): 178S-201S. doi:10.1378/chest.07-1360.

2. W. D. Travis, et al. "International association for the study of lung cancer/American thoracic society/European respiratory society international multidisciplinary classification of lung adenocarcinoma." Journal of thoracic oncology: official publication of the International Association for the Study of Lung Cancer vol. 6, 2 (2011): 244-85. doi:10.1097/JTO.0b013e318206a221

3. L. G. Collins., C. Haines, R. Perkel& R. E. Enck."Lung cancer: diagnosis and management." American family physician vol. 75, 1 (2007): 56-63.

4. K. Yu, C. Zhang, G. Berry, et al. "Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features." Nat Commun 7, 12474 (2016), doi: 10.1038/ncomms12474

5. D. Bazazeh and R. Shubair, "Comparative study of machine learning algorithms for breast cancer detection and diagnosis," 2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA), Ras Al Khaimah, 2016, pp. 1-4, doi: 10.1109/ICEDSA.2016.7818560.

6. E.D. Michie, D.J. Spiegelhalter, and C.C. Taylor, "Machine Learning, Neural and Statistical Classification," Proceeding, 1994.

7. Muthazhagan B, Ravi T, Rajinigirinath D.An enhanced computer-assisted lung cancer detection method using content-based image retrieval and data miningtechniques. Journal of Ambient Intelligence and Humanized Computing.2020 Jun 2:1-9.

8. Masud M, Sikder N, Nahid AA, Bairagi AK, AlZain MA. A machine learning approach to diagnosing lung and colon cancer using a deep learning-basedclassification framework.Sensors. 2021 Jan;21(3):748.

9. Sajja T, Devarapalli R, Kalluri H. Lung Cancer Detection Based on CT Scan Images by Using Deep Transfer Learning. Traitement du Signal. 2019Oct;36(4):339-44.

10. Tripathi P, Tyagi S, Nath M. A comparative analysis of segmentation techniques for lung cancer detection.Pattern Recognition and Image Analysis. 2019Jan;29(1):167-73.

11. Nasrullah N, Sang J, Alam MS, Mateen M, Cai B, Hu H. Automated lung nodule detection and classification using deep learning combined with multiplestrategies. Sensors. 2019 Jan;19(17):3722.

12. Bhatia S, Sinha Y, Goel L. Lung cancer detection: a deep learning approach. InSoft Computing for Problem Solving 2019 (pp. 699-705). Springer, Singapore

13. Makaju S, Prasad PW, Alsadoon A, Singh AK, Elchouemi A. Lung cancer detection using CT scan images. Procedia Computer Science. 2018 Jan 1;125:107-14.

14. Ali I, Hart GR, Gunabushanam G, Liang Y, Muhammad W, Nartowt B, Kane M, Ma X, Deng J. Lung nodule detection via deep reinforcement learning.Frontiers in oncology. 2018 Apr 16;8:108.

15. W. Ausawalaithong, A. Thirach, S. Marukatat, and T. Wilaiprasitporn, "Automatic Lung Cancer Prediction from Chest X-ray Images Using the Deep Learning Approach," 2018 11th Biomedical Engineering International Conference (BMEiCON), Chiang Mai, 2018, pp. 1-5, doi: 10.1109/BMEiCON.2018.8609997.

16. T. Atsushi, T. Tetsuya, K. Yuka, and F. Hiroshi. (, 2017).“Automated Classification of Lung Cancer Types from Cytological Images Using Deep Convolutional Neural Networks”. BioMed Research International. 2017. 1-6. 10.1155/2017/4067832.

17. [11] W. Rahane, H. Dalvi, Y. Magar, A. Kalane and S. Jondhale, "Lung Cancer Detection Using Image Processing and Machine Learning HealthCare," 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT), Coimbatore, 2018, pp. 1-5, doi: 10.1109/ICCTCT.2018.8551008.

18. M. Šarić, M. Russo, M. Stella and M. Sikora, "CNN-based Method for Lung Cancer Detection in Whole Slide Histopathology Images," 2019 4th International Conference

on Smart and Sustainable Technologies (SpliTech), Split, Croatia, 2019, pp. 1-4, doi: 10.23919/SpliTech.2019.8783041.

19. S. Sasikala, M. Bharathi, B. R. Sowmiya. "Lung CancerDetection and Classification Using Deep CNN." (2019).

20. SRSChakravarthy and H. Rajaguru. "Lung CancerDetection using Probabilistic Neural Network with modified Crow-Search Algorithm." Asian Pacific Journal of Cancer Prevention, 20, 7, 2019, 2159-2166, doi: 10.31557/APJCP.2019.20.7.2159.

21. De Bruijne M. Machine Learning Approaches in Medical Image Analysis:From Detection to Diagnosis. Med Image Anal (2016) 33:94–7.doi: 10.1016/j.media.2016.06.032

22. Jindal A, Aujla GS, Kumar N, Chaudhary R, Obaidat MS, You I. Sedative: Sdn-Enabled Deep Learning Architecture for Network Traffic Control in VehicularCyber-Physical Systems. In: IEEE network (2018). p. 66–73.

23. Nalepa J, Kawulok M. Selecting Training Sets for Support Vector Machines:A Review. ArtifIntell Rev (2019) 52:857–900. doi: 10.1007/s10462-017-9611-1

24. Ganesan N, Venkatesh K, Rama M, Palani AM. Application of NeuralNetworks in Diagnosing Cancer Disease Using Demographic Data.Int JComputAppl (2010) 1(26):76–85. doi: 10.5120/476-783.

25. Kasinathan G, Jayakumar S, Gandomi AH, Ramachandran M, Fong SJ, PatanR. Automated 3-D Lung Tumor Detection and Classification by an ActiveContour Model and Cnn Classifier. Expert SystAppl (2019) 134:112–9.doi:10.1016/j.eswa.2019.05.041

26. Shen D, Wu G, Suk HI. Deep Learning in Medical Image Analysis.Annu RevBiomed Eng (2017) 19:221–48.doi: 10.1146/annurev-bioeng-071516-044442

27. Suzuki K. Overview of Deep Learning in Medical Imaging. RadiolPhysTechnol (2017) 10(3):257–73. doi: 10.1007/s12194-017-0406-5

28. 22. Bharati S, Podder P, Mondal MRH. Hybrid Deep Learning for Detecting LungDiseases From X-Ray Images. Inf Med Unlocked (2020) 20:100391. doi:10.1016/j.imu.2020.100391

29. Tajbakhsh N, Suzuki K. Comparing Two Classes of End-to-End Machine-Learning Models in Lung Nodule Detection and Classification: Mtanns vs.Cnns. Pattern Recognit (2017) 63:476–86. doi: 10.1016/j.patcog.2016.09.029

30. Gu Y, Lu X, Yang L, Zhang B, Yu D, Zhao Y, et al. Automatic Lung NoduleDetection Using a 3d Deep Convolutional Neural Network Combined With aMulti-Scale Prediction Strategy in Chest Cts. ComputBiol Med (2018)103:220–31. doi: 10.1016/j.compbiomed.2018.10.011.

31. Sahu P, Yu D, Dasari M, Hou F, Qin H. A Lightweight Multi-Section Cnn forLung Nodule Classification and Malignancy Estimation. IEEE J BiomedHealth Inf (2018) 23(3):960–8. doi: 10.1109/JBHI.2018.2879834

32. Bansal G, Chamola V, Narang P, Kumar S, Raman S. Deep3dscan: DeepResidual Network and Morphological Descriptor Based Framework for LungCancer Classification and 3d Segmentation. IET Image Process (2020) 14(7):1240–7. doi: 10.1049/iet-ipr.2019.1164

33. Jothi G, Inbarani HH. Soft Set Based Feature Selection Approach for LungCancer Images.ArXivPreprint ArXiv (2012) 1212.5391.

34. Anthimopoulos M, Christodoulidis S, Ebner L, Christe A, Mougiakakou S.Lung Pattern Classification for Interstitial Lung Diseases Using a DeepConvolutional Neural Network. IEEE Trans Med Imaging (2016) 35(5):1207–16. doi: 10.1109/TMI.2016.2535865

35. Kasinathan G, Jayakumar S. Cloud-Based Lung Tumor Detection and StageClassification Using Deep Learning Techniques. BioMed Res Int (2022)2022:1–17.doi: 10.1155/2022/4185835

36. Jakimovski G, Davcev D. Using Double Convolution Neural Network for LungCancer Stage Detection. ApplSci (2019) 9(3):427. doi: 10.3390/app9030427

37. Yu H, Zhou Z, Wang Q. Deep Learning Assisted Predict of Lung Cancer onComputed Tomography Images Using the Adaptive Hierarchical HeuristicMathematical Model. In: . IEEE Access (2020). p. 86400–10.

38. T. Sowmiya, M. Gopi, M. New Begin L.Thomas Robinson "Optimization of Lung Cancer using Modern data mining techniques." International Journal of Engineering Research.

39. Ada[1], Rajneet Kaur[2] "Early Detection and Prediction of Lung Cancer Survival using Neural Network Classifier", (IJAIEM)Volume 2, Issue 6, June 2013

40. DasuVaman Ravi Prasad,"Lung cancer detection using image processing techniques", International journal of latest trends in engineering and technology.(2013)

41. S Vishukumar K. Patela and PavanShrivastavab, "Lung A Cancer Classification Using Image Processing", International Journal of Engineering and Innovative Technology Volume 2, Issue 3, September 2012.

42. Fatma Taher1,*, Naoufel Werghi1, Hussain Al-Ahmad1, Rachid Sammouda2, "Lung Cancer Detection Using Artificial Neural Network and Fuzzy Clustering Methods," American Journal of Biomedical Engineering 2012, 2(3): 136-142.