DETECTION OF LUNG CANCER USING IMAGE PROCESSING TECHNIQUES

A. Jahnavi* B. Siva Krishna* D. Lokeshwari*

D. Adithya Kishore Teja* Dr. B. Rajasekhar**

*UG Students, Electronics and Communication Engineering, Seshadri Rao Gudlavalleru Engineering College

* Professor, Electronics and Communication Engineering, Seshadri Rao Gudlavalleru Engineering College

Abstract

Lung cancer remains one of the most life-threatening diseases worldwide, with early detection playing a crucial role in improving patient survival rates. Computed Tomography (CT) scans are among the most effective imaging techniques for detecting lung cancer, yet manual interpretation by doctors can be challenging. To address this issue, computer-aided diagnosis (CAD) systems have been developed to assist in identifying cancerous cells more accurately. This research evaluates various CAD techniques based on image processing and machine learning, analyzing their performance, limitations, and drawbacks. The study systematically reviews different detection approaches, compares their accuracy levels, and highlights areas for improvement. To enhance diagnostic precision, an improved model is proposed that integrates advanced methodologies such as discrete wavelet transform (DWT),

Haralick feature extraction, fuzzy clustering, convolutional neural networks (CNN), and probabilistic neural networks (PNN). The primary objective of this study is to enhance lung cancer detection accuracy, contributing towards more effective and reliable early- stage diagnosis.

Keywords: Lung Cancer Detection, Computer-Aided Diagnosis (CAD), Medical Image Processing, Computed Tomography (CT) Scans, Machine Learning, Deep Learning, Discrete Wavelet Transform (DWT), Haralick Feature Extraction, Fuzzy Clustering, Convolutional Neural Networks (CNN), Probabilistic Neural Networks (PNN), Early Diagnosis, Segmentation, Pattern Recognition.

INTRODUCTION

Lung cancer is one of the leading causes of cancer-related deaths worldwide. Its detection is particularly challenging because symptoms often do not appear until the disease has reached an advanced stage. This late diagnosis significantly reduces the chances of successful treatment, making early detection crucial in lowering the mortality rate. Timely identification and intervention can substantially improve survival rates, emphasizing the need for accurate and reliable diagnostic techniques.

Among various imaging techniques**,** Computed Tomography (CT) scans are considered the most effective for lung cancer diagnosis. CT imaging provides detailed cross-sectional images of lung tissues, enabling the identification of both suspected and unsuspected nodules. Despite its

advantages, there are challenges in interpreting CT scan images. The variation in intensity levels, differences in tissue structures, and the potential for misjudgment by radiologists may lead to difficulty in accurately distinguishing cancerous from non-cancerous regions. Such limitations highlight the need for additional tools to support medical professionals in decision- making.

In recent years, Computer-Aided Diagnosis (CAD) systems have emerged as a valuable supplementary tool for assisting doctors and radiologists in detecting lung cancer with greater precision. CAD techniques integrate advanced image processing and machine learning algorithms to enhance detection accuracy and reduce human errors. Various CAD-based approaches have been developed to improve the classification and segmentation of lung cancer nodules.

However, many existing systems still suffer from limited detection accuracy, requiring further refinements to achieve more reliable results.

This research focuses on studying recent lung cancer detection systems that utilize CT scan imaging. By analyzing their methodologies, performance, and limitations, we aim to identify the most effective techniques currently available. Furthermore, based on this analysis, we propose an improved model that integrates advanced image processing methods and machine learning techniques to enhance accuracy, bringing detection performance closer to 100%. Our study aims to contribute to the ongoing efforts in developing more efficient and precise diagnostic tools for early lung cancer detection, ultimately improving patient outcomes.

LITERATURE SURVEY

Lung cancer detection has been extensively studied by researchers who have proposed and implemented various image processing and machine learning approaches to improve accuracy. These studies aim to enhance the detection and classification of cancerous nodules while addressing the limitations of traditional diagnostic methods. Below is a detailed analysis of some significant contributions to the field

[1]. Aggarwal, Furquan, and Kalra proposed a method to differentiate lung cancer nodules from normal lung structures by analyzing geometrical, statistical, and gray-level features. Their classification approach relied on Linear Discriminant Analysis (LDA) along with an optimal thresholding technique for segmentation.

The system demonstrated an overall accuracy of eighty-four percent, with a sensitivity of ninety-seven point one-four percent and specificity of fifty-three point three-three percent. Despite detecting cancer nodules, this model has critical shortcomings, including the absence of advanced machine learning techniques for classification and the utilization of basic segmentation methods. Given these limitations, incorporating its methodology into a new system is unlikely to lead to significant improvements in detection performance.

[2].Jin, Zhang, and Jin developed a Computer-Aided Diagnosis (CAD) system that employed Convolutional Neural Networks (CNNs) for lung cancer detection. The system achieved an accuracy of eighty- four point six percent, with a sensitivity of eighty-two point five percent and a specificity of eighty-six point seven percent. A key advantage of this model was the

application of a circular filter during the Region of Interest (ROI) extraction phase, reducing computational costs for both the training and recognition steps. However, despite optimizing efficiency, the system's accuracy remains suboptimal, requiring further improvements for better detection performance.

[3].Sangamithraa and Govindaraju introduced a segmentation and classification framework that employed an unsupervised learning algorithm based on K-means clustering to group pixel datasets with similar characteristics.

The classification process was carried out using a Back Propagation Neural Network (BPNN). The model extracted key features, including entropy, correlation, homogeneity, Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index (SSIM), using the Gray-Level Co-occurrence Matrix (GLCM) technique. The proposed system demonstrated an accuracy of approximately ninety point seven percent, surpassing the performance of earlier models. Additionally, it incorporated a median filter during the preprocessing phase to remove noise, an approach that could be valuable in future models for improving detection accuracy.

[4.]Roy, Sirohi, and Patle designed a lung cancer detection system incorporating a Fuzzy Inference System (FIS) in conjunction with an Active Contour Model. The methodology included contrast enhancement through gray transformation, followed by image binarization, with segmentation achieved using an active contour model. The classifier was trained using extracted features such as area, mean, entropy, correlation, major axis length, and minor axis length. The system achieved an accuracy rate of ninety- fourpoint one-two percent. However, a primary drawback of this model is its inability to classify detected nodules as benign or malignant, a limitation that future research can address.

[5].Ignatious and Joseph introduced a lung cancer detection system based on watershed segmentation. The system utilized a Gabor filter during the preprocessing stage to enhance image quality and was compared with models that used neural fuzzy segmentation and region-growing techniques. The proposed method exhibited an accuracy of approximately ninety point one percent, outperforming models that relied on alternative segmentation techniques. A notable advantage of this model was its implementation of marker-controlled watershed segmentation, which effectively addressed the issue of over-segmentation.

However, the model still exhibited certain limitations, including the exclusion of preprocessing techniques such as noise removal and image smoothing, both of which could contribute to higher detection accuracy. Additionally, the model does not classify detected nodules as benign or malignant, an aspect that could be improved by incorporating advanced classification algorithms.

[6].Gonzalez and Ponomaryov developed a system capable of distinguishing between benign and malignant lung cancer nodules. Their method utilized prior knowledge and Hounsfield Unit (HU) values to determine the Region of Interest (ROI). The system extracted shape-based features such as area, eccentricity, circularity, and fractal dimension, alongside textural features including mean, variance, energy, entropy, skewness, contrast, and smoothness. A Support Vector Machine (SVM) classifier was employed for classification. A significant advantage of this model was its

ability to categorize nodules as either benign or malignant.

However, its reliance on prior knowledge of the ROI limits its adaptability across different datasets. Despite this, the classification technique using SVM could serve as a valuable addition to future lung cancer detection frameworks.

PROPOSED MODEL

The primary goal of this project is to develop a reliable method for detecting lung cancer at its early stages with high accuracy. Lung cancer continues to be one of the most significant contributors to cancer-related fatalities worldwide.

A major challenge in combating this disease is that it often goes undetected until it reaches an advanced stage, reducing the chances of successful treatment. Therefore, early and precise detection is crucial, as it can significantly enhance the effectiveness of medical interventions and improve patient survival rates.

By utilizing advanced diagnostic techniques, this project aims to assist healthcare professionals in identifying lung cancer at an early stage, ultimately leading to better patient outcomes and increased life expectancy.

Flow diagram

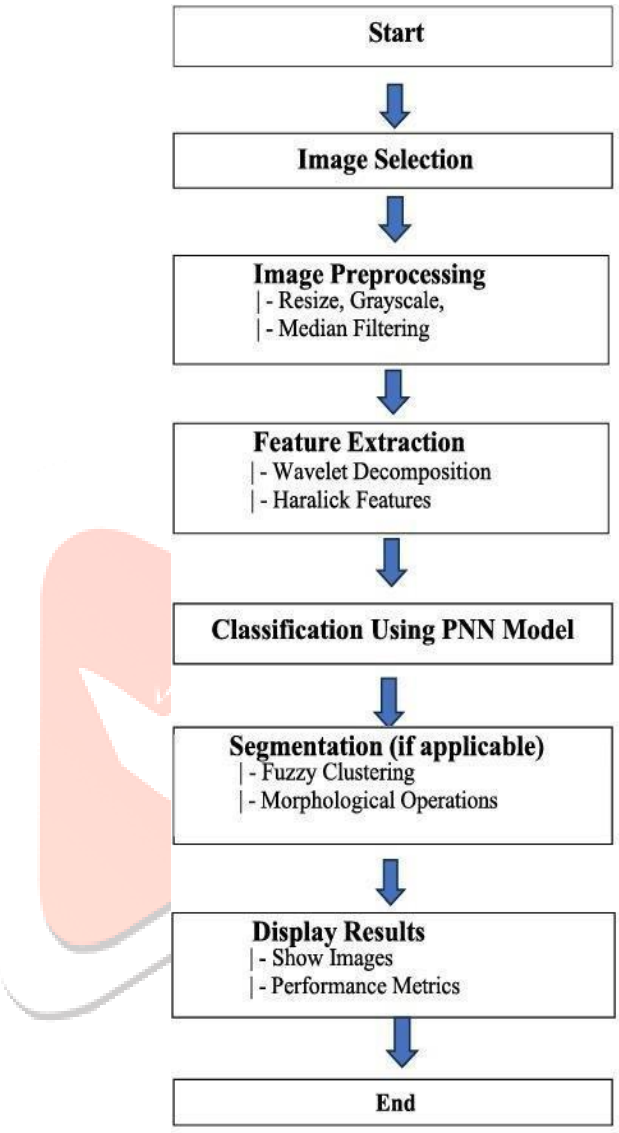


Fig 1: **Flow diagram for initialization for lung cancer detection**

The process begins with system initialization for lung cancer detection. This step involves setting up the required computational environment, selecting necessary algorithms, and preparing the system for image processing. It marks the beginning of the workflow.

Image Selection

At this stage, a medical image is chosen from the Robo Flow universe dataset for analysis. The image is typically a Computed Tomography (CT) scan or an X-ray, which provides a detailed view of lung structures. The selection of high-quality images is crucial for accurate diagnosis, as poor-quality images may lead to incorrect predictions.

Image Preprocessing

Before analysis, the image undergoes preprocessing to enhance quality and remove any distortions. This step ensures that the image is in an optimal format for feature extraction.

Resizing & Grayscale Conversion:

The image is resized to a standardized dimension to ensure consistency across all input images. Conversion to grayscale simplifies processing by reducing computational complexity while retaining important details.

Median Filtering:

This technique is used to remove unwanted noise from the image, such as speckles or grainy distortions. It works by replacing each pixel value with the median value of its surrounding pixels, ensuring smoothness while preserving important edges.

Feature Extraction

This step involves analyzing the image and extracting significant characteristics that help differentiate normal and abnormal lung structures.

Wavelet Decomposition:

This technique breaks down the image into multiple frequency components, allowing finer details to be analyzed at different scales. It helps in identifying abnormalities that may not be visible in the original image.

Haralick Features:

Derived from the Gray-Level Co-occurrence Matrix (GLCM), these features capture important textural properties of the image, such as contrast, correlation, and uniformity. These statistical measures help in distinguishing between healthy lung tissue and cancerous nodules.

Classification Using PNN Model

The Probabilistic Neural Network (PNN) is employed to classify the extracted features and determine whether the lung region contains cancerous nodules.

The PNN model is advantageous due to its fast processing speed and ability to handle noisy datasets effectively. It works by comparing the extracted features against a predefined set of known characteristics to predict whether the given lung scan exhibits signs of cancer. The classification result provides insights into the probability of cancer presence based on statistical comparisons.

Segmentation

Segmentation is performed when additional isolation of cancerous regions is required. This step helps in precisely identifying the affected areas within the lung image.

Fuzzy Clustering:

This method groups pixels into clusters based on their intensity and similarity. Instead of making a rigid classification, fuzzy clustering assigns each pixel a probability of belonging to a specific region, which helps in distinguishing tumors from normal lung tissues.

Morphological Operations:

These are image processing techniques that refine the segmented regions by removing noise and enhancing the shape of detected objects. It involves operations like dilation, erosion, opening, and closing to improve segmentation accuracy and ensure that detected cancerous areas are well-defined.

Display Results

Once classification and segmentation are completed, the system presents the final output.

Show Images:

The processed lung images are displayed, with highlighted regions indicating potential cancerous areas. This visualization helps medical professionals in assessing the results effectively.

Performance Metrics:

The system calculates key performance indicators such as accuracy, sensitivity, specificity, precision, and F1-score to evaluate how well the model performs. These metrics provide insights into the reliability of the detection system.

IMPLEMENTATION

For the implementation of the proposed lung cancer detection model, real CT scan images of patients were obtained from the Lung Image Database Consortium (LIDC) archive the LIDC Dataset is obtained from the Kaggle Website. This database serves as a benchmark dataset for lung cancer research and is widely used for developing, training, and evaluating computer-aided diagnostic (CAD) systems for early lung cancer detection. The LIDC database comprises several cases, contributed by seven academic centers and eight medical imaging companies, providing a diverse and reliable dataset for research. The images in this database are stored in DICOM (Digital Imaging and Communications in Medicine) format**,** which is the standard format used in medical imaging, with each image having a resolution of 256 × 256 pixels to ensure high-quality medical imaging.

Since DICOM images are complex to process due to their specialized metadata structure and encoding, they were converted into grayscale JPEG format using MicroDicom software. MicroDicom enables researchers to open and analyze DICOM CT scan images while also allowing conversion into more commonly used formats such as JPEG, making them more accessible for processing in MATLAB. After converting the images, the proposed model was developed and implemented using MATLAB R2013a**,** which is widely used for research, data analysis, and algorithm development in medical imaging.

The implementation in MATLAB involved two major components: detection and feature extraction, both of which were

carried out using MATLAB's Image Processing Toolbox**.** The feature extraction process focused on wavelet decomposition- based texture analysis and Gray-Level Co- occurrence Matrix (GLCM) features**,** which help in identifying potential lung nodules by analyzing the texture and structure of lung tissues. Once the features were extracted, classification was performed using MATLAB's Classification Learner Toolbox, which simplifies the process of training machine learning models. The classification step involved categorizing detected lung nodules into different stages, such as **normal, benign, or malignant**, based on the extracted features.

To enhance the reliability of the classification model and prevent overfitting**, a** 5-fold cross-validation technique was applied. This technique ensures that the model generalizes well to unseen data by dividing the dataset into five subsets, using four subsets for training while reserving one for validation in each iteration. The process is repeated five times so that each subset is used for validation once, reducing the risk of overfitting and improving overall performance.

For training the model, 16 DICOM images from the LIDC dataset were used**,** containing multiple lung nodules. To validate the performance of the trained model**,** 5 additional images containing a total of 15 lung nodules were used as test data. This helped in assessing the accuracy and effectiveness of the proposed system in detecting and classifying lung cancer cases.

Despite the successful implementation of the proposed model, several challenges were encountered during the process. One of the major challenges was the large size of the LIDC database**,** which is 124 GB**.** Downloading and managing such a vast dataset was time-consuming and required efficient data handling strategies to optimize storage and processing. Another challenge was dealing with the complex cancer annotations provided in XML format. The LIDC dataset includes detailed XML files containing lung nodule annotations made by multiple radiologists. Extracting relevant information from these XML files was challenging, as they contained multiple annotation layers and metadata descriptions that required careful parsing and interpretation. Overcoming these challenges was crucial for successfully implementing the lung cancer detection model and ensuring its efficiency in real-world applications.

RESULTS

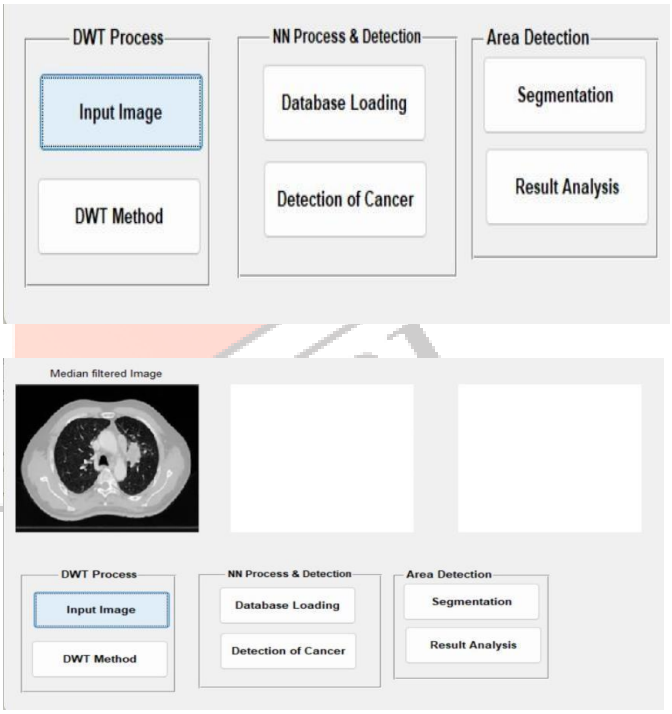
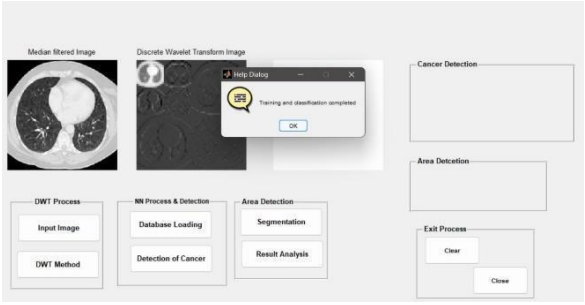


Fig2: Selection input image

The image highlights the "Input Image" button within the "DWT Process" module, indicating the user is at the initial stage of selecting a medical image for analysis. This step is crucial for initiating the image



processing pipeline, likely for cancer detection

Fig3: Training the PNN model with CT scan lung images and classification are done using extracted features

This image depicts a GUI for a medical image analysis application, likely for lung cancer detection. It showcases a workflow involving Discrete Wavelet Transform (DWT), neural network processing, and area detection. A "Help Dialog" box confirms the completion of training and classification, suggesting the use of a Probabilistic Neural Network (PNN) model, as indicated in the caption

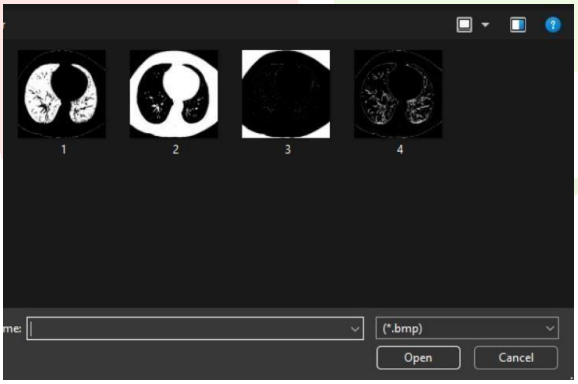


Fig 4: Clusters of segmentation

This image depicts a file selection window on a computer screen. It showcases four lung scan images, labeled 1 through 4, likely the results of a segmentation process. The window includes standard file dialog options like "Open" and "Cancel," suggesting the user is choosing a specific scan file.

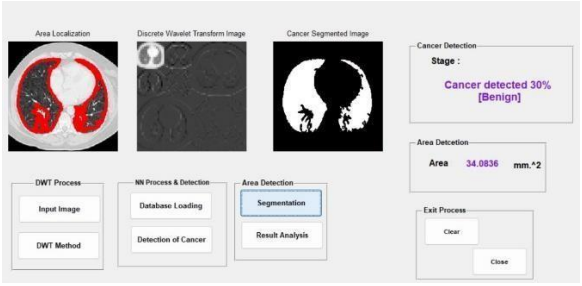


Fig 5: Detection of tumor and performing segmentation using fuzzy clustering method

his image depicts a software interface for lung cancer detection, utilizing fuzzy clustering. It shows a lung scan with highlighted areas, alongside analysis steps like wavelet transform and segmentation. The software reports a "Cancer detected 30% [Benign]" result and an area measurement of 34.0836 mm^2.

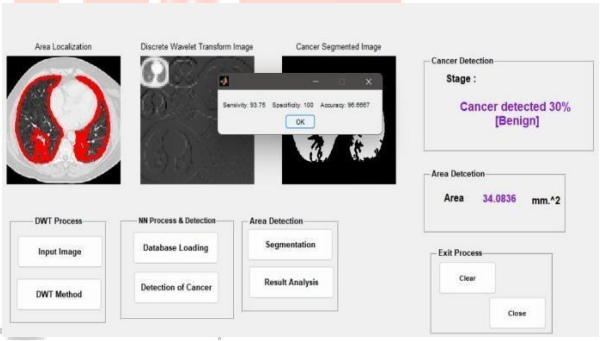


Fig 5: Display of sensitivity, accuracy, specificity

This image shows a software interface for lung cancer detection, displaying results and process steps. It features a lung scan with highlighted areas, alongside analysis metrics like sensitivity (92.75%), specificity (100%), and accuracy (98.66%). The interface also shows a "Cancer detected 30% [Benign]" result and an area measurement of 34.0836 mm^2, indicating the software's diagnostic output.

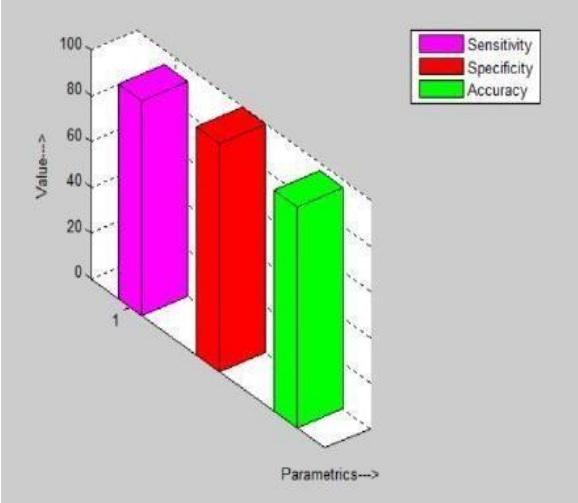


Fig6: Graphical Representation of Sensitivity, accuracy, specificity

This image presents a 3D bar graph comparing Sensitivity, Specificity, and Accuracy, likely of a diagnostic test or model. Sensitivity is highest, followed by Specificity, with Accuracy being the lowest among the three metrics. The graph provides a visual representation of the model's performance in these three key areas

CONCLUSION

The primary goal of this project is to achieve early and accurate detection of lung cancer, as it remains one of the leading causes of cancer-related deaths worldwide. Timely identification of cancerous nodules is crucial for improving patient survival rates. Existing models have limitations in terms of accuracy and fail to classify the severity of detected nodules. To address these shortcomings, the proposed system integrates DWT decomposition and Harlick features are extracted using convolutional neural networks (CNN) to trained the probabilistic neural networks (PNN), for classification into Malignant or Benign or normal.

This model demonstrates significant improvement, achieving 96% classification accuracy, which is higher than the current best-performing system, with an additional category (healthy lung)

However, a limitation of the proposed model is its inability to classify cancer into different stages such as Stage I, II, III, and IV. Future improvements could focus on integrating a stage-wise classification system to enhance diagnostic precision. Additionally, optimizing the preprocessing techniques and minimizing false object detections could further enhance accuracy, making the model more reliable for real-world medical applications.

REFERENCES

[1] Gindi,A. M., Al Attiatalla, T. A., & Sami,
M.M. (2014) "A Comparative Study for Comparing Two Feature Extraction Methods and Two Classifiers in Classification of Earlystage Lung Cancer Diagnosis of chest x- ray images." Journal of American Science, 10(6): 13-22.

[2] Suzuki, K., Kusumoto, M., Watanabe, S. I., Tsuchiya, R., & Asamura, H. (2006) "Radiologic classification of small adenocarcinoma of the lung: radiologic-pathologic correlation and its prognostic impact," The Annals of Thoracic Surgery. 81(2): 413-419.

[3] Xiuhua,G., Tao, S., & Zhigang, L.(2011) "Prediction Models for Malignant Pulmonary Nodules Based-on Texture Features of CT Image." In Theory and Applications of CT Imaging and Analysis. DOI: 10.5772/14766.

[4] Aggarwal, T., Furqan, A., & Kalra, K. (2015) "Feature extraction and LDA based classification of lung nodules in chest CT scan images." 2015 International Conference

On Advances In Computing, Communications And Informatics (ICACCI), DOI: 10.1109/ICACCI.2015.7275773.

[5] 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). DOI: 10.1109/ISCID.2016.1053.

[6] Sangamithraa, P., & Govindaraju, S. (2016) "Lung tumour detection and classification using EK-Mean clustering." 2016 International Conference On Wireless Communications, Signal Processing And Networking (Wispnet). DOI: 10.1109/WiSPNET.2016.7566533.

[7] Roy, T., Sirohi, N., & Patle, A. (2015) "Classification of lung image and nodule detection using fuzzy inference system." International Conference On Computing, Communication & Automation. DOI: 10.1109/CCAA.2015.7148560.

[8] Ignatious, S., & Joseph, R. (2015) "Computer aided lung cancer detection system." 2015 Global Conference On Communication Technologies (GCCT), DOI: 10.1109/GCCT.2015.7342723.

[9] Rendon-Gonzalez, E., & Ponomaryov, V. (2016) "Automatic Lung nodule segmentation and classification in CT images based on SVM." 2016 9Th International Kharkiv Symposium On Physics And Engineering Of Microwaves, Millimeter And Submillimeter Waves (MSMW). DOI: 10.1109/MSMW.2016.7537995.

[10] Miah, M.B.A., & Yousuf, M.A. (2015) "Detection of lung cancer from CT image using image processing and neural network." 2015 International Conference on Electrical Engineering and Information Communication Technology (ICEEICT): 1-

[11] The 6th International Conference on Smart Computing and Communications Edited by Jimson Mathew, Ashutosh K. Singh https://www.sciencedirect.com/science/article/pii/S1877050917327801