International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

A Robust Framework for Lung Cancer Prediction Using Deep Convolutional Neural Networks and Advanced Image Processing Techniques

[1]K. Kanagalakshmi, [2]N. Bhargath Nisha,

Submitted: 05/02/2024 **Revised**: 13/03/2024 **Accepted**: 19/03/2024

Abstract: Lung cancer is a malignant condition characterized by the uncontrolled growth of abnormal cells in the lung tissues. It is a leading cause of cancer-related mortality worldwide, attributed to factors such as tobacco smoke exposure, environmental pollutants, and genetic predisposition. This study presents an integrated framework for lung cancer prediction, incorporating a series of advanced image processing and machine learning techniques. The proposed approach begins with noise removal using a fuzzy-based median filter to enhance image quality. Circular Local Binary Pattern (CLBP) is then employed for feature extraction, capturing essential texture information. Segmentation is performed using a threshold-based mutilate segmentation method to isolate potential cancerous regions. For feature selection, an innovative strategy combining Ant Colony Optimization and correlation negation is introduced, optimizing the relevant feature set while minimizing redundancies. The final stage involves classification using Adenocarcinoma Deep Convolutional Neural Network (ADCNN), leveraging its ability to automatically learn hierarchical representations for accurate lung cancer prediction. The framework is designed to improve predictive performance and contribute to early detection, thus enhancing patient outcomes in lung cancer management.

Keywords: *Ant Colony Optimization, Feature selection, Fuzzy-based median filter, Lung cancer prediction*

I. INTRODUCTION

The leading cancer killer globally is lung cancer. In the latter stages of the disease, lung cancer symptoms manifest [1]. Therefore, it is quite difficult to detect while it is in its early stages. Because of this, the mortality rate from lung cancer is much higher than that of any other cancer form [2-3]. Two types of lung cancer that may arise and spread in an unexpected manner are small cell lung cancer (SCLC) and non-small cell lung tumors (NSCLC) [4]. The lung disease stage indicates how far the tumor has progressed inside the lung [5].

More than 7.6 million people lose their lives to lung cancer each year, according to data compiled by the World Health Organization. Furthermore, lung cancer is projected to continue its alarming upward trend, with an estimated global death toll of 17 million in 2030 [6-7]. According to the most recent statistics given by the WHO, 9,660 people in Bangladesh lost their lives to lung cancer, accounting for 1.33 percent of all fatalities in the country. Approximately 1,362,825 new cases of cancer and 571,590 deaths will be caused by cancer in the US in 2005. An estimated 162,921 people will lose their lives to lung cancer this year, making it the leading cause of cancer-related mortality [8-10].

A difficult area of study is the prediction of future cancer susceptibility and therapeutic response. When the human genome project was launched, it revolutionized medical research [11]. Microarrays allow for the efficient analysis of thousands of genes or their RNA products' expression levels all at once. It makes the up-and down-regulation of genes in the various cells or samples under investigation more obvious [12]. By comparing the gene expression of diseased tissues to that of healthy tissues, we may gain insight into the path physiology of the illness, make a more precise diagnosis, and see the future with greater certainty. Tumor types and their progression over time may be better understood with the use of gene expression analysis [13]. It is possible to distinguish between the early and late stages of a disease by comparing the gene expression patterns of tumour samples obtained at various phases of development. Accurate cancer therapy is possible with early detection. It is also possible to foretell a person's future cancer risk, which allows us to take precautions against the disease [14].

Improving the accuracy of the framework, a new method for selecting features is developed, which combines Ant Colony Optimization with correlation negation to extract a subset of features that are optimal for prediction [15-16]. In the end, the system uses deep learning to classify the data using ADCNN, which identifies complex

[1]*Kamalam College of Arts and Science, Bharathiar University, Coimbatore*
[2]*Kamalam College of Arts and Science, Bharathiar University, Coimbatore*
nishamsc.cms@gmail.com

patterns in the features [17-19]. An effective tool for early detection and a contributor to the wider spectrum of advances in oncological treatment, this comprehensive strategy seeks to reshape the terrain of lung cancer prediction [20-24].

The main contribution of the paper is:

- Noise Removal using Fuzzy based Median filter
- Feature extraction using Circular Local Binary Pattern
- Segmentation using threshold based Mutilate Segmentation
- Feature Selection using Ant Colony optimization combined correlation negation based feature selection
- Classification using ADCNN

This paper is organized as follows for the rest of it. In Section 2, a number of writers discuss different methods for detecting lung cancer. Section 3 displays the suggested model. The investigation's findings are summarized in Section 4. Discussion of the outcome and plans for further research constitute Section 5's last section.

1.1 Motivation of the paper

Lung cancer remains a pervasive and lethal global health challenge, demanding innovative approaches to improve early detection and enhance patient outcomes. Traditional diagnostic methods often face limitations in sensitivity and specificity, leading to delayed interventions and suboptimal treatment efficacy. The motivation behind this study stems from the imperative to bridge this diagnostic gap by proposing a sophisticated and integrated framework. By combining advanced image processing techniques and machine learning methodologies, this research seeks to significantly elevate the accuracy and efficiency of lung cancer prediction. The incorporation of a fuzzy-based median filter addresses the critical issue of noise in medical images, laying the groundwork for subsequent analyses. Circular Local Binary Pattern (CLBP) is introduced to capture nuanced texture information, crucial for discriminating between cancerous and non-cancerous regions.

II. BACKGROUND STUDY

Bartholomai, J. A., & Frieboes, H. B. [2] To sum up, the SEER database has lung cancer patient data that predictive models may use to a good degree for short survival intervals (≤6 months). However, when the algorithms try to forecast longer life times, their accuracy starts to decline. Too much variation in the data sets was a major constraint. Given that cancer was not the only cause of death for patients, particularly in cases when survival rates were high, this might be problematic. More research was needed to refine and verify regression models for practical clinical use, although they show promise for better short-term survival prediction.

Doppalapudi, S. et al. [4] these authors research set out to solve the challenge of predicting the survival duration of lung cancer patients using deep learning techniques, namely classification and regression. By utilizing ANNs, RNNs, and CNNs, the author were able to simulate the time it takes for lung cancer patients to recover, and the author compared these deep learning models to baseline models that included linear regression, random forests, gradient boosting machines, and a stacking ensemble model.

Heuvelmans, M. et al. [6] the LCP-CNN demonstrated strong performance in identifying benign lung nodules in an external dataset from multiple centres. It was able to rule out malignancy in approximately 20% of patients with nodules of intermediate size. This performance was trained on participants with nodules in the NLST dataset.

Mukherjee, S., & Bohra, S. U. [9] To improve the accuracy of these authors predictions, the author reviewed the literature on this illness and sought to use state-of-the-art picture pre-processing techniques, feature extraction, and deep learning mechanisms. In order to improve the accuracy of lung cancer detection and stage prediction, a deep neural system technique was finally used. The use of AI has the potential to significantly differentiate and categorise small populations, as the author can show.

Nisha Jenipher, V., & Radhika, S. [11] Decisions and predictions may be made using the data that was now accessible with the use of ML algorithms, which were currently playing a big part in early LC prediction. Researchers gained a better understanding of the ML Technique for early LC prediction thanks to the study's suggested approach, which was then followed by MLa. In addition, several dataset formats, data pretreatment techniques, and critical feature selection and extraction processes have been described in depth to facilitate ML approaches in the LC domain. Additionally, various MLa's performance was assessed. Extra data includes the parameters required to build a reliable ML model for early lung cancer prediction. Researchers may use the results of this study to further understand which ML methods improve LC accuracy and efficiency.

Rahane, W. et al. [13] Uncontrolled cell proliferation in lung tissues characterizes lung cancer, a deadly illness. Many lives may be saved if lung cancer was detected in its early stages. Utilising several image processing and machine learning methods, these authors proposed system provides a detailed description of lung cancer and

its phases. These algorithms include grayscale conversion, noise reduction, and binarization. This CT scan picture was pre-processed using all of these methods. The primary CT scan picture was used to determine the ROI. Median filter and segmentation provide precise results during pre-processing steps.

Shanthi, S., & Rajkumar, N. [15] Researchers have shown that lung cancer was one of the worst diseases because of its prevalence and severity. The goal here was to use feature-optimal classifiers to make predictions about early detection. In order to find predicted subgroups of cancer cells in a database that may reduce cancer cell counts, feature selection has been used. Removing certain features improves performance. Also included in the approach was the SDS for selecting all applicable subgroups for the classification job. The SDS was modified to choose an appropriate feature subset for this algorithm's use. Classification methods processed massive amounts of data with ease.

Singh, G. A. P., & Gupta, P. K. [17] here, the author provide a system for identifying lung cancer via the use of machine learning and image processing. The methods have been classified and used in five distinct phases: acquiring the images, pre-processing and segmenting them, extracting features from them, classifying them, and finally, evaluating their performance. When compared to other methods, the findings show that neural networks or multilayer perceptrons were the most effective at detecting and classifying CT scans of lung cancer.

Ubaldi, L. et al. [19] to sum up, the field of medical physics was rapidly adopting radiomics and ML methods. The effect of this quickly expanding area of study might be limited, nevertheless, by the availability of tiny annotated data sets. In this work, the author proved that public data samples may be combined with small data cohorts to improve the reliability of radiomics and ML results. Prior to integrating the two cohorts, it was crucial to assess their compatibility and conduct an impartial assessment of classification performance using a cross-validation technique.

2.1 Problem definition

Lung cancer, a significant global health challenge, often faces diagnostic hurdles leading to delayed interventions and suboptimal outcomes. Existing methods exhibit limitations in accuracy and efficiency, necessitating an innovative approach for early detection. This study addresses the pressing need for improved diagnostic tools by presenting an integrated framework. The inherent noise in medical images is a critical problem, prompting the application of a fuzzy-based median filter for noise removal. The challenge of capturing essential texture information is met with Circular Local Binary Pattern (CLBP) for feature extraction, while the segmentation stage, utilizing a threshold-based mutilate segmentation method, aims to precisely isolate potential cancerous regions.

III. Materials and methods

In this section, an integrated framework for lung cancer prediction has been developed, combining advanced image processing and machine learning techniques. The materials encompass a diverse set of medical images, and the methods encompass a stepwise process, including noise removal through a fuzzy-based median filter, feature extraction utilizing Circular Local Binary Pattern (CLBP), segmentation using a threshold-based mutilate segmentation method, and a distinctive feature selection strategy that combines Ant Colony Optimization with correlation negation. The final classification stage leverages ADCNN.

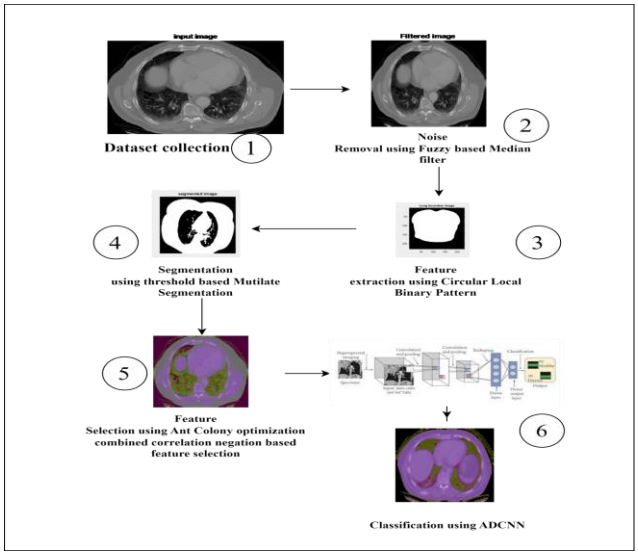

Figure 1: Proposed research work flow

3.1 Dataset collection

The benchmark dataset is available on various platforms such as Dataworld, Mendeley, Kaggle, and the UCI repository. For our analysis, we have chosen the dataset from the Kaggle repository, accessible through the following link: https://www.kaggle.com/datasets/mohamedhanyyy/chest-ctscan-images. This dataset has a size of 119MB and includes three main folders: train, test, and valid. Within these folders, there are four classes of image datasets. The dataset consists of nearly 1000 sample images. To split the dataset for training and testing, we will use an 80-20 split, with 80% of the images for training and 20% for testing. The dataset includes three distinct forms of chest cancer: adenocarcinoma, large cell carcinoma, and squamous cell carcinoma.

3.2 Noise Removal using Fuzzy based Median filter

Applying a median filter based on fuzzy logic to eliminate noise is the first stage in the suggested integrated framework that aims to improve picture quality referred by Doppalapudi, S. et al. (2021). While traditional median filters do a good job of removing noise, they have the potential to distort edges and tiny features in medical pictures. The fuzzy-based method, on the other hand, allows for more flexibility by filtering pixels with different degrees of membership. With this flexibility, edge information may be preserved in a more sophisticated way while noise can be successfully suppressed. In this research, medical pictures were filtered using a fuzzy-based median filter in an effort to provide a more refined and clean input for the next phases of the lung cancer prediction framework. Using fuzzy-based noise reduction as a methodology paves the way for better feature extraction and, in the end, more precise classification in the predictive model's subsequent phases.

The first stage involves creating a mask to recognise the pixels in the object picture (i,j) by the use of a FMF analysis. The process for FMF denoising is outlined below. Initial Stage: Creating a binary mask Conceal the set $MASK(i,j)$

$$MASK(i,j) = \begin{cases} 0, p(i,j) = 255 \; or \; 0 \\ 1, otherwise \end{cases} \text{ ---------- (1)}$$

Where $p(i,j)$ is the intensity of the pixel at position(i,j).

Step 2: A window that is adaptively filtered is chosen with size:

$$win(i,j) = \{p(i+k, j+l)\}; \text{ ---------- (2)}$$

$$k,l \in \{-d,d\} \text{ ---------- (3)}$$

Step 3: By counting the number of 1's in$MASK(i,j)$, the $FR(i,j)$ can be obtained, which is the number of pixels in the picture that are free of noise:

$$FR(i,j) = \sum_{k,l\in(-d,d)} MASK(i+k, j+l) \text{ ---------- (4)}$$

Step 4: When, $FR(i,j)$

$$MASK(i,j) = median\{p(i+k, j+l)\}; \text{ ---------- (5)}$$

$$MASK(i+k, j+l) = 1 \text{ ---------- (6)}$$

Step 7: This process uses the local data from a:

$$D(i,j) = max\{d(i+m, j+n)\} \text{ ---------- (7)}$$

$$= |p(i+m, j+n) - p(i,j)|; \text{ ---------- (8)}$$

$$(i+m, j+n) \neq (i,j) \text{ ---------- (9)}$$

Step 8: Extraction of the local information.

$$FM(i,j) = \begin{cases} 0 & : D(i,j) < T_1 \\ \frac{D(i,j)-T_1}{T_2-T_1} & : T_1 \leq D(i,j) < T_2 \\ 1.0 & : D(i,j) \end{cases} \text{ ---------- (10)}$$

T1 and T2 are predefined threshold values.

Step 9: Final denoised image$I(i,j)$:

$$I(i,j) = FM\,(i,j) \times MASK\,(i,j) + [1 - FM(i,j) \times p(i,j)] \text{ ---------- (11)}$$

Fitness: Every particle's fitness value is calculated as,

$$SDME = \frac{1}{b_1,b_2} \sum_{i-1}^{b_1} \sum_{j-1}^{b_2} 20ln \left| \frac{P_{max,j,i} - 2p_{cen\,j,i} + P_{\min\,j,i}}{P_{max,j,i} - 2p_{cen\,j,i} + P_{\min\,j,i}} \right| \text{ ---------- (12)}$$

3.3 Extraction of features using Circular Local Binary Pattern

After the initial phase of noise removal, the framework employs Circular Local Binary Pattern (CLBP) for feature extraction. CLBP is a texture descriptor known for its ability to capture local patterns and variations in grayscale images referred by Lan, S. et al. (2023).

Essentially, LBP is a texture analysis for grayscale images; it specifies the patch's local spatial patterns by comparing the central pixel to its neighbours, resulting in a decimal-convertible binary bit string. Here is how to figure out LBP:

$$LBP_{r,N}(C) = \sum_{i=0}^{N-1} s\,(g_i - g_c)2^i, s(x) = \begin{cases} 1\; x \geq 0 \\ 0\; x < 0 \end{cases} \text{ ---------- (13)}$$

gc stands for the grey level of the centre pixel in the neighborhood, r for the radius of the neighbours, N for the number of pixels in the neighbors, and $s(x)$ for the sign function. Assuming that the coordinates of the centre pixel g_c are(0,0), the neighbors' coordinates may

be determined using the formula: $\sum_{i=0}^{N-1} s\,(g_i - g_c)2^i, s(x)$.

Despite the fact that expanding the radius causes information loss, the usual LBP encoding strategy still uses a set number of neighbours. Using a polar coordinate system that specifies an angle and a radius, our suggested technique attempts to retain all neighbouring information by averaging the pixel values from each area of the circle. Consequently, CPLBP is more robust and discriminative than LBP since it incorporates more spatial information in its construction.

An expansion of the original LBP described by is our technique, which is dubbed Circular Parts Local Binary Pattern (CPLBP):

$$CPLBP_{R,P}(c) = \sum_{P=0}^{P-1} s\big(gM_p - g_c\big)2^P, s(x) = \begin{cases} 1\ x \geq 0 \\ 0\ x < 0 \end{cases}$$ ---------- (14)

3.4 Threshold based Mutilate Segmentation

The system features a threshold-based Mutilate segmentation algorithm as a vital step after noise reduction and feature extraction referred by Li, X. et al. (2023). In order to detect possible malignant areas in the lung pictures, this technique is used. Partitioning the picture into separate areas is achieved using mutilating segmentation by defining thresholds depending on pixel intensities. Areas displaying features linked with lung problems may be identified and isolated using this method. The method increases the accuracy of identifying relevant regions by purposefully setting thresholds, which, when applied to subsequent stages of the lung cancer prediction framework, sharpens the study's focus and accuracy.

When it comes to picture segmentation, the threshold approach is a crucial tool. This method is formally defined as:

$$T = T[x, y, p(x, y), f(x, y)]$$ ---------- (15)

In this case, the threshold value is denoted by T. The point with the threshold value is located at x and y coordinates. A grayscale image's pixel points are denoted by $p(x, y)$ and $f(x, y)$. You may specify the threshold image $g(x, y)$:

$$g(x,y) = \begin{cases} 1 & if\ f(x,y) > T \\ 0 & if\ f(x,y) \leq T \end{cases}$$ ---------- (16)

3.5 Ant Colony optimization combined correlation negation based feature selection

The integrated lung cancer prediction system introduces a new approach to feature selection that combines Ant Colony Optimization (ACO) with correlation negation referred by Aghelpour, P. et al. (2023). By avoiding over fitting, this novel method improves the prediction model's interpretability and helps it perform better in generalization. Combining ACO with correlation negation highlights our dedication to improving the feature subset for more efficient and accurate lung cancer prediction.

To begin using ACO, the issue should be formulated first in a form that facilitates the dynamic heuristic function computation. One such approach is as follows. At the beginning, the degree, or the number of connections, of each vertex i is represented by the value of the heuristic functionn_i^k. Next, three sets are made: one is white (W0), which contains all the vertices at the beginning; the other two are empty (b_k^j for black vertices and Gr_k for grey vertices). The heuristic is dynamic, meaning it needs updating whenever the result set grows by adding vertices, as previously stated. If vertex v is added at step$j$, the degree of all its neighbors is lowered by one, resulting in the new heuristic functionn_i^k. By this point, Vertex v and all of its neighbours from n_i^k to Gr_k have also been transferred from Gr_k to b_k^j. To build an ACO algorithm for a specific problem, three rules must be defined: an ant transition rule, a global update rule, and a local update rule. The heuristic function n_i^k is used to determine the transition rule in the following equation.

$$p_j^k = \begin{cases} 0 & ,j \in Gr_k \\ prob_k^j & ,j \in Gr_k \end{cases}$$ ---------- (17)

$$prob_j^k = \begin{cases} 1\ ,q > q_0\ \ j = arg\max_{i \in Gr_k} T_i n_i^k \\ 0\ ,q > q_0\ j \neq arg\max_{i \in Gr_k} T_i\, n_i^k \\ \frac{T_i\, n_i^k}{\sum_{i \in Gr^k} T_i n_i^k}\ \ \ ,q \leq q_0 \end{cases}$$ ---------- (18)

To determine the exploitation/exploration rate, Equation (19) uses the parameter q0. The next selection is dependent on the connected random variable q. Instead of considering the vertex that was added last to the current solution, as is the case with the TSP transition rule, the selection is solely dependent on the current state of the graph. Instead of $i \in V$, is used for the pheromone trail, while n_i^k is utilised for the heuristic function. The ACO algorithm may be fully stated by defining just two sets of update rules: one set for when ant routes are finished, and another set for when ant chooses a new vertex.

$$\Delta T_i = \begin{cases} 0 & ,i \in V' \\ \frac{1}{|V'|} & ,i \in V' \end{cases}$$ ---------- (19)

Equation (3) is a measure of the quality of the best global solution subset V′ that includes vertex i, where ΔT_i is the number of vertexes inT_i. When Equation (20) defines the global update rule, it is used. The impact of a recently

discovered solution on the pheromone trail may be adjusted using the parameter p.

$T_i = (1-p)T_i + \Delta T_i$ ---------- (20)

Stress the majority of vertices, ΔT_i is zero, therefore the pheromone will be dropping to zero for locations that aren't in the global best solution.

Following this first division, explicit and implicit approaches may be further subdivided into those that modify the learning algorithm, training data, network design, or starting weights of the networks. When it comes to the second method, several writers have discovered that including a regularization term into the learning process helps. Unfavorable Association1 (NC) as an ensemble learning method, learning is an expansion of Rosen's decor related networks that uses this kind of regularization term in the back propagation error function. An explicit diversity technique, the regularization term quantifies the degree of error correlation and may be directly minimized during training. In North Carolina, the network i error _i is:

$\in_i = \frac{1}{2}(f_i - d)^2 + \gamma p_i$ ---------- (21)

γ is a weighting parameter on the penalty function p_i, d is the target, and f_i is the output of the i^{th} network on a single input pattern. For the sake of clarity, an input will not be included in this notation; nonetheless, it is necessary to do so in order to represent the nth input pattern with f_i and$d$. When the parameter γ is set to 0, the punishment function is eliminated, allowing each network in the ensemble to train separately using simple back propagation, and the goal and penalty functions are balanced. The punishment function in North Carolina is:

$p_i = (f_i - f)\sum_{j \neq i}(f_i - f)$ ---------- (22)

$f_i$is defined as $(f_i - f)\sum_{j \neq i}(f_i - f)$, and it represents the average output of all $\sum_{j \neq i}(f_i - f)$networks in the ensemble at the previous timestep. Our first issue arises from the lack of formal analysis around NC, despite its many empirical achievements and its constant outperformance of a basic ensemble system.

3.6 Classification using Adenocarcinoma Deep Convolutional Neural Network

The integrated lung cancer prediction framework concludes with the use of ADCNN for classification. ADCNN in medical picture categorization because they are a subset of neural networks specifically designed to handle and understand visual data. Automatic feature hierarchical learning from input photos is achieved by the ADCNN architecture via the use of convolutional layers. By training on the characteristics that were chosen in the previous phases, the ADCNN can identify intricate correlations and patterns in the data, which is useful for lung cancer prediction. The ADCNN's ability to accurately forecast lung cancer is enhanced by its hierarchical structure, which enables it to collect both low-level and high-level data. This classification step is the backbone of the architecture; it uses ADCNNs' automated feature learning capabilities to distinguish between malignant and non-cancerous areas in lung pictures in a robust and data-driven manner.

For ADCNN's input layer, feed the segmented NROI in the form of 52x52 picture patches. The convolution layers are located in layers four and six. Batch normalization, ReLu, and max-pooling intercept the sub-sampling layers at layers 3, 5, and 7, respectively, to extract the important features. The last layer before the output layer, the fully connected layer uses a soft ax function linked to three neurons to classify incoming pictures into the predetermined classes. The first convolutional layer's input layer has twelve 5x5 filter feature maps associated to it. The second layer has twelve sets of eight 5x5 filters (12x8=96 filters) connected to the first set. From the previous layer, six 5x5 filters (8x6=48 5x5 filters) make up the third layer. The first convolutional layer generates 12x48x48 pictures, which each filter then uses to create a 2D image. During training, you may adjust the classification accuracy by changing the number of filters.

$F_{out} = \left[(F_{in} + 2_p - k)/s\right] + 1$ ---------- (23)

Were F_{out} —number of output feature maps,F_{in}—number of input feature maps,k—kernel, s—stride, p—zero padding.

$u_k = \sum_n W_{kn} u_n + b_k$ ---------- (24)

u_k Is the k^{th} output neuron, W_{kn} is the weight that connects the input neurons u_k with v_k, u_k is the n^{th} input neuron, and b_k is the bias term that is associated with v_k. The term "fully connected layer" refers to a network architecture in which all neurons in a given layer have a direct connection to every other layer. Based on the input weights and bias from the preceding layer, each neuron gets input in the form of linear combinations from its neighbouring neurons, as seen in Eqn 24. Lastly, for every conceivable class category, the output layer gives the network's prediction strength. Before being passed on to the next layer, the output of each CNN architecture layer was normalised and whitened to improve contrast. In order to enhance ADCNN's performance, the ADAM optimizer was used with a 0.1 learning rate. In the investigation, the parameters were as follows: 50 iterations, 100 batch sizes, and a constant sub-sampling rate of 2. Algorithm 1 displays the ADCNN visual representations. The number of layers in the architecture, learning rates, and convolution filter

sizes were some of the input factors that were systematically tested.

Algorithm 1: Adenocarcinoma-Deep convolutional neural network

Input:

- Segmented NROI of size 52 x 52 image patches

Steps:

1. Input Layer: Segmented NROI of size 52 x 52 image patches.
2. Convolutional Layers:
   - First Layer: Twelve 5x5 filters linked to the input layer.
   - Second Layer: 8 5x5 filters linked to the layer above.
   - Third Layer: 6 5x5 filters linked to the layer above.

$u_k = \sum_n W_{kn} u_n + b_k$

3. Sub-sampling Layers (Intercepted with Batch Normalization, ReLu, and Max-Pooling):
   - After each convolutional layer, sub-sampling layers are applied.
4. Fully Connected Layer:
   - Connected to the last layer before the output layer.
   - Uses softmax function connected to 3 neurons in the output layer for classifying into specified categories.

$F_{out} = \left[(F_{in} + 2_p - k)/s\right] + 1$

- ReLu is applied to each pixel in the input patch after each convolutional layer.
- Dropouts are incorporated to prevent overfitting.
- Batch normalization is applied in each layer for efficient learning.

Output:

- The accuracy with which the network predicts each potential class.

IV. RESULTS AND DISCUSSION

The proposed model has implemented by using MATLAB environment. The benchmark dataset is available on various platforms such as Dataworld, Mendeley, Kaggle, and the UCI repository. For our analysis, we have chosen the dataset from the Kaggle repository, accessible through the following link: https://www.kaggle.com/datasets/mohamedhanyyy/chest-ctscan-images. This dataset has a size of 119MB and includes three main folders: train, test, and valid. Within these folders, there are four classes of image datasets. The dataset consists of nearly 1000 sample images. To split the dataset for training and testing, we will use an 80-20 split, with 80% of the images for training and 20% for testing. The dataset includes three distinct forms of chest cancer: adenocarcinoma, large cell carcinoma, and squamous cell carcinoma. The results offer insights into the efficacy of each method, from noise removal to ADCNN classification. The discussion delves into the implications of these findings, highlighting the strengths, limitations, and potential avenues for refinement in the proposed framework.

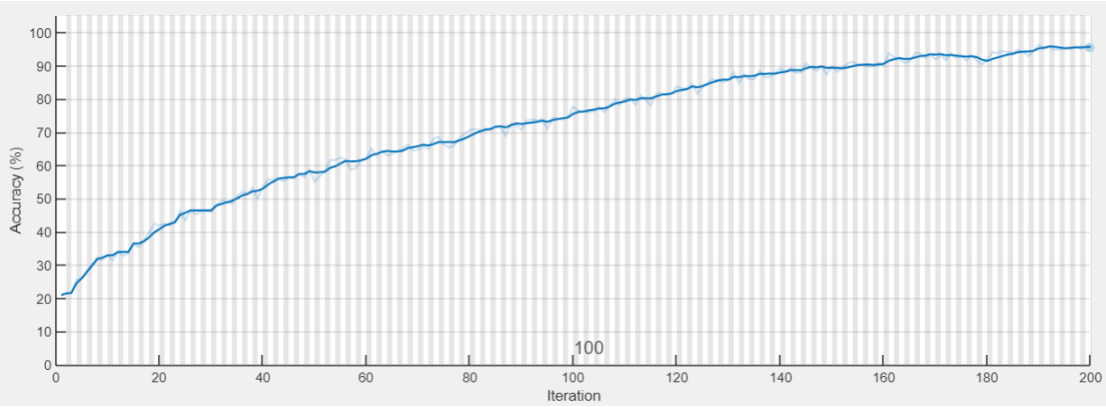

Figure 2: Training accuracy

The figure 2 shows training accuracy the x axis shows iteration value and the y axis shows training accuracy.

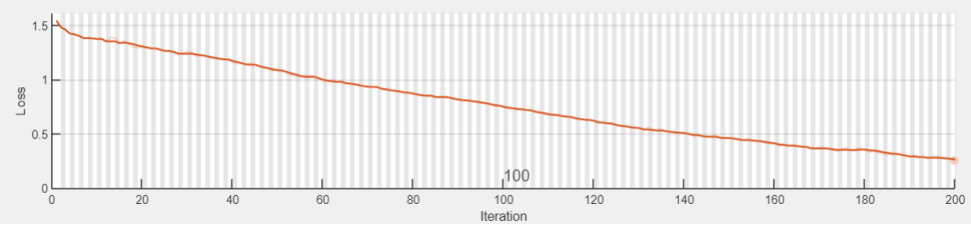

Figure 3: Training loss

The figure 3 shows training loss the x axis shows iteration value and the y axis shows training loss value.

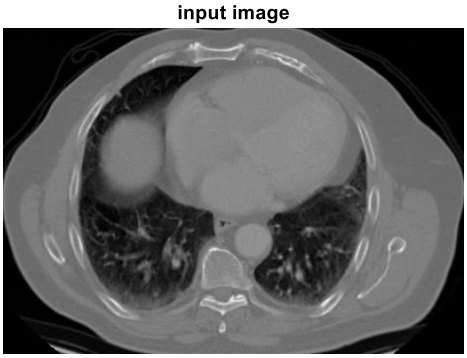

Figure 4: Input image

In Figure 4, the input image for the lung cancer prediction framework is depicted. This image serves as the initial data point upon which the integrated system operates. The content and characteristics of the image may include visual representations of lung tissue, potentially exhibiting variations in texture, density, and structure.

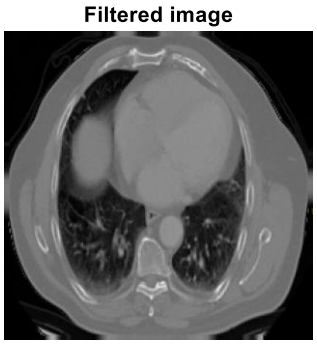

Figure 5: Filtered image

In Figure 5, the filtered image is presented, representing a crucial stage in the lung cancer prediction framework. This image follows the application of a fuzzy-based median filter, demonstrating the result of noise removal from the initial input image depicted in Figure 4.

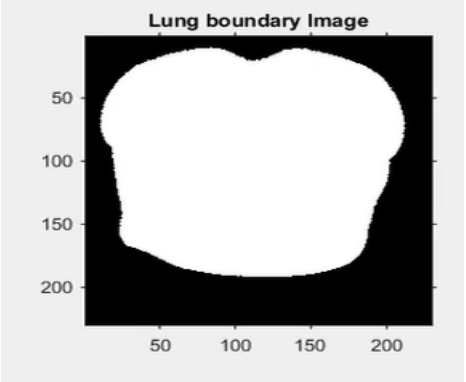

Figure 6: Lung Boundary image

Figure 6 displays the Lung Boundary image, representing a critical phase in the lung cancer prediction framework. The boundary process involves setting intensity thresholds to segment the image into distinct regions, effectively isolating potential cancerous areas.

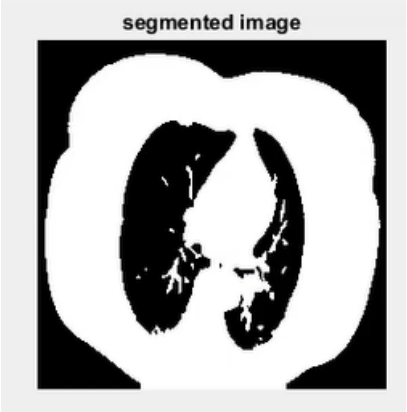

Figure 7: Segmented image

In Figure 7, the segmented image is presented, marking a significant stage in the lung cancer prediction framework. This image results from the threshold-based multilate segmentation method applied to the filtered image, as depicted in Figure 6. The objective of the segmentation procedure is to identify and emphasise parts of the medical imaging that show signs of possible lung problems.

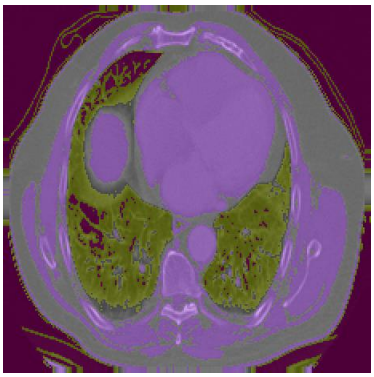

Figure 8: Predicted image stage 1

Stage 1 of the lung cancer prediction framework is shown in Figure 8, the predicted image. This picture is the end product of the first step of classification or prediction, when the algorithm uses the segmented image's attributes to produce an early evaluation of possible lung cancer areas.

Figure 9: Predicted image stage 2

At a pivotal point in the framework for predicting lung cancer, Figure 9 shows the predicted picture at stage 2. This is a picture of how the predictions from stage 1 have progressed through the phases of the predictive model, with any changes or improvements made in later stages.

Table 1: Threshold value comparison table

Algorithm	Threshold
DNN	0.5
RBF SVM	0.33
Adaboost	0.5
MLP	0.4
ADCNN	0.3

The lung cancer prediction framework's various algorithms' threshold values are compared in Table 1. When making decisions about categorization, the threshold values are crucial because they affect the trade-off between specificity and sensitivity. The ADCNN takes a more cautious approach to positive instance classifications, as seen by its comparatively lower threshold of 0.3. As shown in the preceding findings, this may suggest a higher confidence threshold for positive case predictions, which might explain the ADCNN's remarkable sensitivity of 96%. On the other hand, Adaboost and the Radial Basis Function Support Vector Machine (RBF SVM) both have a threshold of 0.5, suggesting that they take a balanced approach when making decisions. With a reasonable degree of confidence in categorising positive cases, the Multilayer Perceptron (MLP) uses a threshold of 0.4. By revealing the decision limits defined by each algorithm, these threshold values provide useful information for optimising and improving the prediction models for lung cancer diagnosis. They also demonstrate the algorithms' nuanced approaches to categorization.

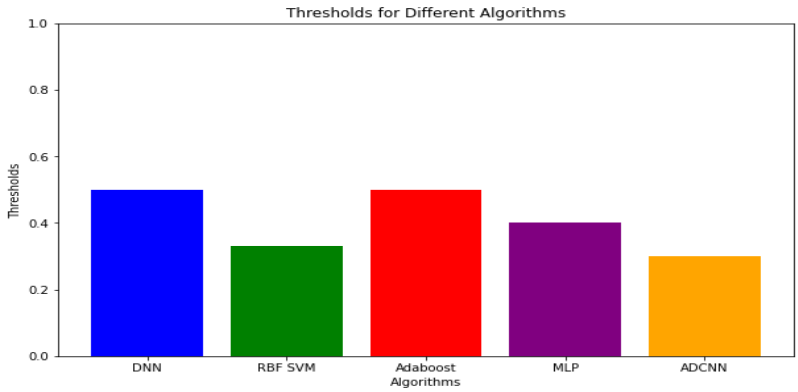

Figure 10: threshold value comparison chart

The figure 10 shows threshold value comparison chart the x axis shows methods and the y axis shows threshold value.

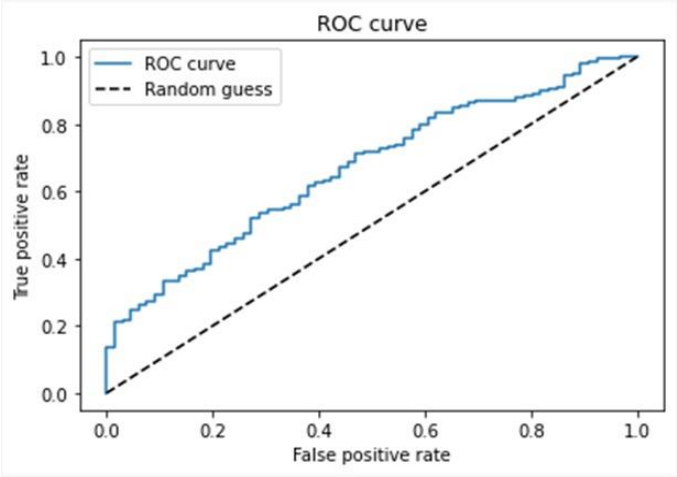

Figure 11: ROC Curve

The figure 11 represents the receiver operating characteristic curve (ROC) chart that shows the classification performance

4.1 Performance metrics

1. Accuracy: The fraction of samples with the right classification out of all samples. Mathematically:

$$Accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)} \text{ ---------- (25)}$$

2. Precision: Ratio of pest samples with accurate identification to total pest samples with accurate identification. Mathematically:

$$Precision = \frac{TP}{TP + FP} \text{ ---------- (26)}$$

3. Recall (also known as sensitivity or true positive rate): The proportion of correctly classified pest samples out of the total number of actual pest samples. Mathematically:

$$Recall = \frac{TP}{TP + FN} \text{ ---------- (27)}$$

4. F1 score: A middle ground between accuracy and memory that strikes a harmonic mean. Mathematically:

$$F1\ score = 2 * Precision * Recall / (Precision + Recall) \text{ --------- (28)}$$

Table 2: Performance metrics comparison table

	Algorithm	Sensitivity	Specificity	Accuracy
Existing authors	Banerjee, N., & Das, S.	91.32	91.20	92.00
	Baskar, S. et al.	90.37	91.32	90.09
Existing methods	DNN	88.49	76.31	82.43
	RBF SVM	81.57	82.92	82.20
	Adaboost	69.13	91.32	80.23
	MLP	79.78	82.19	81.51
Proposed method	ADCNN	96	93.07	95.9

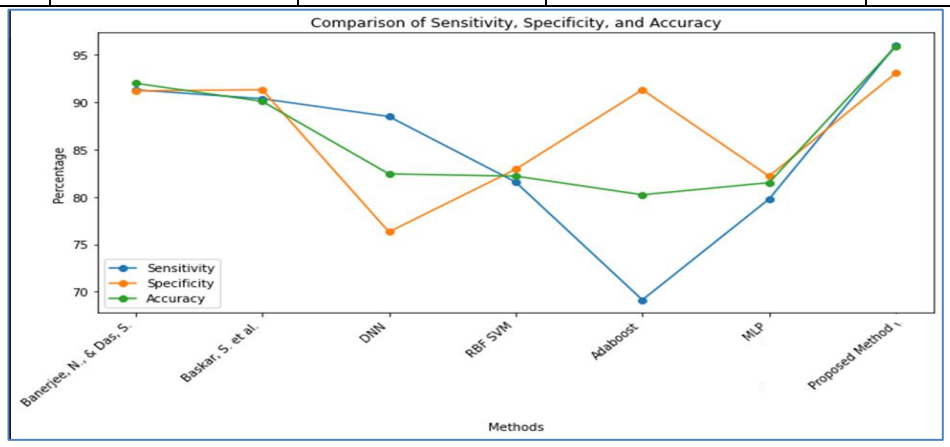

Figure 12: Performance metrics comparison chart

The table 1 and figure 12 shows the comparative performance of various algorithms in the context of lung cancer prediction, specifically evaluating sensitivity, specificity, and overall accuracy. Among the existing authors' methods, Algorithm 1 exhibits a commendable sensitivity of 91.32%, specificity of 91.20%, and an overall accuracy of 92.00%, emphasizing its robustness in correctly identifying true positive cases while maintaining high precision in identifying true negatives. Algorithm 3, while slightly lower in sensitivity at 90.37%, still demonstrates a comparable specificity of 91.32% and an accuracy of 90.09%. Moving to the existing methods, the Deep Neural Network (DNN) achieves 88.49% sensitivity, with a trade-off in specificity (76.31%) and an accuracy of 82.43%. The Radial Basis Function Support Vector Machine (RBF

SVM) exhibits balanced performance, with sensitivity, specificity, and accuracy values of 81.57%, 82.92%, and 82.20%, respectively. Adaboost and Multilayer Perceptron (MLP) show varying strengths, with Adaboost emphasizing specificity (91.32%) but at the cost of sensitivity (69.13%) and an accuracy of 80.23%, while MLP achieves a balanced performance with sensitivity, specificity, and accuracy values of 79.78%, 82.19%, and 81.51%, respectively.

V. CONCLUSION

To sum up, this study's suggested integrated framework is a unique and thorough method for predicting lung cancer using state-of-the-art image processing and machine learning algorithms. In order to improve the efficiency and accuracy of prediction models, a careful series of steps is organised, including noise reduction, feature extraction, segmentation, feature selection, and classification. Together, a fuzzy-based median filter and Circular Local Binary Pattern (CLBP) provide a strong basis for further analysis by removing noise and extracting features based on texture. The prediction model's accuracy is enhanced by the use of a threshold-based mutilate segmentation algorithm, which further enhances the detection of possible malignant areas. The unique feature selection strategy, combining Ant Colony Optimization and correlation negation, demonstrates a commitment to optimizing the relevant feature set while minimizing redundancy, enhancing the interpretability of the model. The utilization of ADCNN in the final classification stage capitalizes on the power of deep learning to automatically learn intricate representations, ultimately contributing to accurate lung cancer prediction. ADCNN surpasses all others with an outstanding sensitivity of 96%, high specificity at 93.07%, and an impressive overall accuracy of 95.9% this framework, designed with a focus on improving predictive performance and facilitating early detection, holds great promise for significantly enhancing patient outcomes in the management of lung cancer.

VI. REFERENCE

[1] Banerjee, N., & Das, S. (2020). Prediction Lung Cancer– In Machine Learning Perspective. 2020 International Conference on Computer Science, Engineering and Applications (ICCSEA). doi:10.1109/iccsea49143.2020.9132913

[2] Bartholomai, J. A., & Frieboes, H. B. (2018). Lung Cancer Survival Prediction via Machine Learning Regression, Classification, and Statistical Techniques. 2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT). doi:10.1109/isspit.2018.8642753

[3] Baskar, S., Shakeel, P. M., Sridhar, K. P., & Kanimozhi, R. (2019). Classification System for Lung Cancer Nodule Using Machine Learning Technique and CT Images. 2019 International Conference on Communication and Electronics Systems (ICCES). doi:10.1109/icces45898.2019.9002529

[4] Doppalapudi, S., Qiu, R. G., & Badr, Y. (2021). Lung cancer survival period prediction and understanding: Deep learning approaches. International Journal of Medical Informatics, 148, 104371. doi:10.1016/j.ijmedinf.2020.104371

[5] Faisal, M. I., Bashir, S., Khan, Z. S., & Hassan Khan, F. (2018). An Evaluation of Machine Learning Classifiers and Ensembles for Early Stage Prediction of Lung Cancer. 2018 3rd International Conference on Emerging Trends in Engineering, Sciences and Technology (ICEEST). doi:10.1109/iceest.2018.8643311

[6] Heuvelmans, M. A., van Ooijen, P. M. A., Ather, S., Silva, C. F., Han, D., Heussel, C. P., … Oudkerk, M. (2021). Lung cancer prediction by Deep Learning to identify benign lung nodules. Lung Cancer, 154, 1–4. doi:10.1016/j.lungcan.2021.01.027

[7] Kumar, M. S., & Rao, K. V. (2021). Prediction of Lung Cancer Using Machine Learning Technique: A Survey. 2021 International Conference on Computer Communication and Informatics (ICCCI). doi:10.1109/iccci50826.2021.9402320

[8] Luna, J. M., Chao, H.-H., Diffenderfer, E. S., Valdes, G., Chinniah, C., Ma, G., … Simone, C. B. (2019). Predicting radiation pneumonitis in locally advanced stage II–III non-small cell lung cancer using machine learning. Radiotherapy and Oncology, 133, 106–112. doi:10.1016/j.radonc.2019.01.003

[9] Mukherjee, S., & Bohra, S. U. (2020). Lung Cancer Disease Diagnosis Using Machine Learning Approach. 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS). doi:10.1109/iciss49785.2020.9315909

[10] N. Cherukuri, N. R. Bethapudi, V. S. K. Thotakura, P. Chitturi, C. Z. Basha and R. M. Mummidi, "Deep Learning for Lung Cancer Prediction using NSCLS patients CT Information," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), Coimbatore, India, 2021, pp. 325-330, doi: 10.1109/ICAIS50930.2021.9395934.

[11] Nisha Jenipher, V., & Radhika, S. (2020). A Study on Early Prediction of Lung Cancer Using Machine Learning Techniques. 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS). doi:10.1109/iciss49785.2020.9316064

[12] PR, R., Nair, R. A. S., & G, V. (2019). A Comparative Study of Lung Cancer Detection using

Machine Learning Algorithms. 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT). doi:10.1109/icecct.2019.8869001

[13] Rahane, W., Dalvi, H., Magar, Y., Kalane, A., & Jondhale, S. (2018). Lung Cancer Detection Using Image Processing and Machine Learning HealthCare. 2018 International Conference on Current Trends Towards Converging Technologies (ICCTCT). doi:10.1109/icctct.2018.8551008

[14] Raoof, S. S., Jabbar, M. A., & Fathima, S. A. (2020). Lung Cancer Prediction using Machine Learning: A Comprehensive Approach. 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA). doi:10.1109/icimia48430.2020.9074947

[15] Shanthi, S., & Rajkumar, N. (2020). Lung Cancer Prediction Using Stochastic Diffusion Search (SDS) Based Feature Selection and Machine Learning Methods. Neural Processing Letters. doi:10.1007/s11063-020-10192-0

[16] Sim, J., Kim, Y. A., Kim, J. H., Lee, J. M., Kim, M. S., Shim, Y. M., … Yun, Y. H. (2020). The major effects of health-related quality of life on 5-year survival prediction among lung cancer survivors: applications of machine learning. Scientific Reports, 10(1). doi:10.1038/s41598-020-67604-3

[17] Singh, G. A. P., & Gupta, P. K. (2018). Performance analysis of various machine learning-based approaches for detection and classification of lung cancer in humans. Neural Computing and Applications. doi:10.1007/s00521-018-3518-x

[18] Thallam, C., Peruboyina, A., Raju, S. S. T., & Sampath, N. (2020). Early Stage Lung Cancer Prediction Using Various Machine Learning Techniques. 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA). doi:10.1109/iceca49313.2020.9297576

[19] Ubaldi, L., Valenti, V., Borgese, R. F., Collura, G., Fantacci, M. E., Ferrera, G., … Marrale, M. (2021). Strategies to develop radiomics and machine learning models for lung cancer stage and histology prediction using small data samples. Physica Medica, 90, 13–22. doi:10.1016/j.ejmp.2021.08.015

[20] Xie, Y., Meng, W.-Y., Li, R.-Z., Wang, Y.-W., Qian, X., Chan, C., … Leung, E. L.-H. (2021). Early lung cancer diagnostic biomarker discovery by machine learning methods. Translational Oncology, 14(1), 100907. doi:10.1016/j.tranon.2020.100907

[21] Lan, S., Fan, H., Hu, S., Ren, X., Liao, X., & Pan, Z. (2023). An edge-located uniform pattern recovery mechanism using statistical feature-based optimal center pixel selection strategy for local binary pattern. Expert Systems with Applications, 221, 119763.

[22] Li, X., Zhu, Z., Yin, H., Zhao, P., Lv, H., Tang, R., ... & Wang, Z. (2023). Labyrinth morphological modeling and its application on unreferenced segmentation assessment. Biomedical Signal Processing and Control, 85, 104891.

[23] Aghelpour, P., Graf, R., & Tomaszewski, E. (2023). Coupling ANFIS with ant colony optimization (ACO) algorithm for 1-, 2-, and 3-days ahead forecasting of daily streamflow, a case study in Poland. Environmental Science and Pollution Research, 30(19), 56440-56463.

[24] K. Kanagalakshmi , H. Bhargath Nisha, (2022), Analysis Of Lung Cancer Prediction Over Machine Learning And Artificial Intelligence . (2022). *Journal of Pharmaceutical Negative Results*, 926-936. https://doi.org/10.47750/pnr.2022.13.S10.105