LUNG CANCER PREDICTION USING IMPROVED VGG19 NETWORKS AND ADVANCED IMAGE PROCESSING TECHNIQUES

*Dr. M. Sukanya**

ABSTRACT

Lung cancer remains a leading cause of cancer-related deaths worldwide, necessitating advancements in early detection and accurate diagnosis. This study proposes a robust methodology for lung cancer prediction leveraging an improved VGG19 network in conjunction with advanced image processing techniques. The proposed framework integrates multiple stages to enhance the precision of detection and classification. Initially, noise in the lung imaging data is mitigated using a fuzzy-based median filter, which effectively reduces noise while preserving crucial structural details. This preprocessing step is critical for enhancing the quality of the input data and ensuring reliable feature extraction. Following noise removal, feature extraction is conducted using the Circular Local Binary Pattern (CLBP) method. CLBP captures the intricate textural and spatial information within the lung images, which is essential for distinguishing between benign and malignant regions. The segmented images are then processed through a threshold-based Mutilate Segmentation technique. This segmentation method accurately isolates the regions of interest (ROIs) within the lung images, ensuring that only relevant sections are analyzed in subsequent steps. Finally, the classification stage employs an improved VGG19 network, optimized for lung cancer detection. Modifications to the traditional VGG19 architecture enhance its capability to recognize patterns specific to lung cancer, resulting in higher accuracy and reduced false-positive rates. The integration of these advanced image processing techniques with the improved VGG19 network presents a comprehensive solution for lung cancer prediction. This study underscores the importance of combining sophisticated image processing methods with deep learning models to achieve superior performance in medical image analysis.

Keywords: Advanced image processing, Circular Local Binary Pattern, Improved VGG19, Lung cancer, Mutilate Segmentation

I. INTRODUCTION

Lung cancer is a major global health problem due to the substantial impact it has on patient outcomes, necessitating the urgent development of state-of-the-art prediction models [1]. In order to achieve this goal, we have created an innovative framework that combines machine learning with cutting-edge image processing [2]. Early detection is key in the battle against this dangerous disease, which is why a comprehensive approach was developed to account for the nuances of medical image noise and the capabilities of convolutional neural networks [3]. This study delves into every aspect, from using a fuzzy-based median filter to eliminate noise to utilising CLBP for thorough feature extraction [4]. Our goal is to provide researchers and doctors with a potent new tool for lung cancer prediction by integrating various methods, raising the standard for accuracy, and transforming the field [5].

Since lung cancer is a big public health issue globally, there needs to be a significant shift in the way it is detected [6]. In order to tackle this important problem, our research proposes a state-of-the-art integrated framework that employs state-of-the-art methods for machine learning and image processing [7-8]. When patients are diagnosed early, their treatment outcomes and chances of survival are significantly enhanced [9]. The suggested architecture organises a meticulous sequence of steps, the first of which is the use of a median filter based on fuzzy logic to surgically improve medical photographs and remove noise [10, 11]. From here, we train the model to differentiate between

Department of CSE (Cyber Security)
Karpagam Academy of Higher Education, Coimbatore, Tamil Nadu, India
sukanya.murugesan@kahedu.edu.in
* Corresponding Author

healthy and malignant tissue by extracting features using CLBP, which enables it to detect granular texture characteristics [12]. The next stage is to use a threshold-based multilate segmentation strategy that will aid in the strategic discovery of potentially cancerous sections for in-depth analysis.

This document will be structured in the following manner from now on. Section 2 delves into many methods for detecting lung cancer, as explored by various writers. Finally, in Section 3, we see the suggested model. The findings of the investigation are described in Section 4. Results and future study goals are discussed in Section 5, which finishes the section.

1.1 Motivation of the paper

Delays in treatments and less-than-ideal treatment success are common results of traditional diagnostic procedures' lack of sensitivity and specificity. The urgent need to provide a complex and comprehensive framework to close this diagnostic gap is the driving force behind our research. To tackle the important problem of noise in medical pictures, a fuzzy-based median filter is used, which sets the stage for future analysis. A new method called CLBP has been developed to detect malignant and non-cancerous areas by capturing detailed texture information.

II. BACKGROUND STUDY

Finally, for short survival periods ($\leqslant$ 6 months), prediction models may make effective use of the lung cancer patient data included in the SEER database. Unfortunately, the algorithms' accuracy begins to deteriorate as they attempt to predict longer life spans. One big limitation was the abundance of data set variation. The fact that cancer was not the only killer of patients, especially in circumstances with good survival rates, raises concerns about this. Although regression models show promise for improved short-term survival prediction, more research is necessary to enhance and validate them for practical clinical usage[13].

This study's authors set out to use deep learning methods, namely regression and classification, to the difficult problem of estimating how long lung cancer patients would survive. When tested on an external dataset compiled from several centres, the LCP-CNN proved to be very effective at detecting benign lung nodules. For around 20% of individuals with intermediate-sized nodules, it was able to rule out cancer. The NLST dataset included subjects with nodules for the purpose of training this performance[14].

Because it was so common and deadly, lung cancer was ranked high among the worst illnesses by researchers. Making predictions about early detection using feature-optimal classifiers was the objective here. Feature selection has been used to identify cancer cell subgroups in a database that might lead to a decrease in cancer cell numbers. You can enhance performance by removing certain features. The strategy also made use of the SDS to pick out all relevant subgroups for the categorization task. To make sure this approach was applied to the right subset of features, the SDS was adjusted. Classification algorithms effortlessly handled enormous data sets[15].

2.1 Problem definition

Diagnostic delays in lung cancer, a major public health concern worldwide, can cause treatments to be postponed and results to be less than ideal. A fresh strategy is required for early detection since current approaches are inefficient and inaccurate. Through the presentation of an integrated framework, this research tackles the urgent need for enhanced diagnostic tools.

III. MATERIALS AND METHODS

This section describes the whole method that was utilised to develop a system that may forecast the occurrence of lung cancer by using cutting-edge techniques in machine learning and image processing.

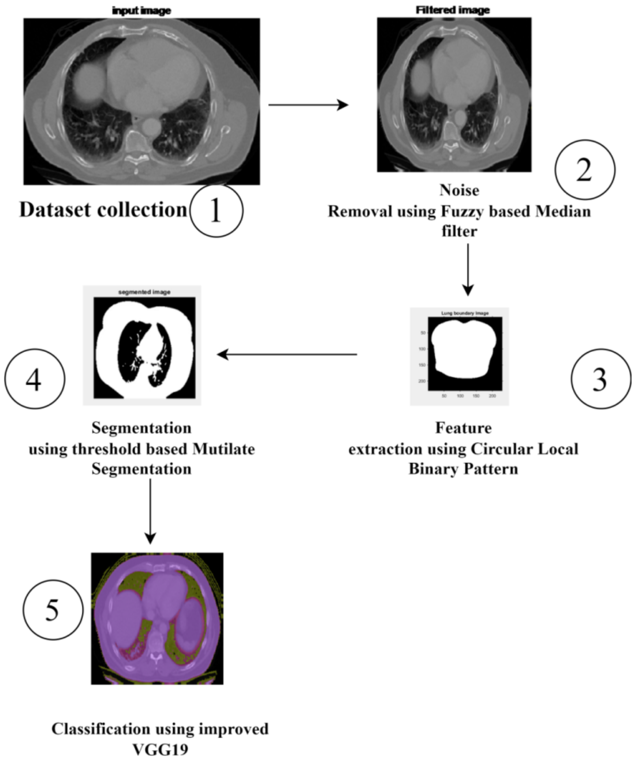
input image
Filtered image
Dataset collection
1
2
Noise
Removal using Fuzzy based Median filter
segmented image
Lung boundary image
4
3
Segmentation
using threshold based Mutilate Segmentation
Feature
extraction using Circular Local Binary Pattern
5
Classification using improved VGG19

Figure 1: Proposed research work flow

3.1 Dataset collection

The dataset was collected from Kaggle website https://www.kaggle.com/datasets/mohamedhanyyy/chest-ctscan-images The information encompasses adenocarcinoma, squamous cell carcinoma, and giant cell carcinoma, which are three separate types of chest cancer. On top of that, regular cells fall into one particular group.

3.2 Noise Removal using Fuzzy based Median filter

The first step in the proposed integrated framework to enhance image quality, is to apply a fuzzy logic-based median filter to remove noise. Conventional median filters are effective in reducing background noise, but they may skew sharp edges and small details in medical images. However, by excluding pixels with varying degrees of membership, the fuzzy-based approach provides more leeway. This adaptability allows for the potential preservation of edge information in a more nuanced manner, along with the effective suppression of noise. In the latter stages of the predictive model, using fuzzy-based noise reduction as a technique allows for greater feature extraction and, ultimately, more accurate classification.

In the first step, we apply a fast moving forward (FMF) analysis to create a mask that can identify the pixels in the object image (i,j). What follows is an explanation of how to do FMF denoising. Step One: Make a binary mask. The set MASK(i,j) must be unknown.

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

Step 2: Adaptively filtered windows are selected based on size

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

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

Step 3: We can get the number of noise-free pixels in the image, FR(i,j), by counting the 1's in MASK(i,j).

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

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

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

3.3 Extraction of features using Circular Local Binary Pattern

When the first round of noise reduction is complete, the framework uses CLBP to extract features. CLBP captures local patterns and changes in grayscale pictures, making it a texture descriptor. For the purpose of lung cancer prediction, CLBP is very helpful in identifying fine-grained textural variations in medical pictures. To make it more resistant to rotation fluctuations, CLBP incorporates circular sampling, unlike typical Local Binary Patterns (LBP). The procedure entails making a histogram of the image's encodings, which are the local texture patterns, after cutting it into circular areas. In order to differentiate between healthy and malignant areas in lung scans, the framework intends to represent unique characteristics by extracting important texture information using CLBP. This is a crucial first step since it provides a wealth of discriminative information for better predictive modelling in the next phases.

In its most basic form, LBP is a grayscale picture texture analysis that describes the patch's local spatial patterns via a decimal-convertible binary bit string produced by comparing the centre pixel to its neighbours. Here is how low back pain is diagnosed:

$$LBP_{r,N}(C) = \sum_{i=0}^{N-1} s\,(g_i - g_c)2^i, s(x) = \begin{cases} 1 \; x \geq 0 \\ 0 \; x < 0 \end{cases} \text{ ----------- (6)}$$

Using the formula: ∑_(i=0)^(N-1)

may be used to compute the neighbours' coordinates, supposing that the centre pixel g_c has coordinates

(0, 0).| s(x) = (g_i - g_c)/2^i

Despite the fact that data loss occurs as the radius is increased, the conventional LBP encoding technique persists in using a constant number of neighbours. Since CPLBP uses more spatial information in its creation, it is more robust and discriminative than LBP.

$$CPLBP_{R,P}(c) = \sum_{P=0}^{P-1} s(gM_p - g_c)2^P, s(x) = \begin{cases} 1 \; x \geq 0 \\ 0 \; x < 0 \end{cases} \text{ -------- (7)}$$

3.4 Threshold based Mutilate Segmentation

Using this method, potential cancerous regions in lung images may be identified. By establishing thresholds based on pixel intensities, multilate segmentation partitions the image into distinct regions. This technique allows for the identification and isolation of areas exhibiting characteristics associated with lung diseases. This approach improves the precision of the research by making it easier to pinpoint important areas by deliberately establishing thresholds; this, in turn, improves the accuracy of the future phases of the lung cancer prediction framework. By zeroing down on potential pathologically significant regions for feature selection and classification, threshold-based Multilate segmentation improves the overall prediction model.

The threshold technique is an essential tool for image segmentation. In formal terms, this approach is described as

$$T = T[x, y, p(x, y), f(x, y)] \text{ -------------------- (8)}$$

Here, T stands for the threshold value. At the given x and y coordinates is where the threshold value is found. Both p(x,y) and f(x,y) represent pixel coordinates in a grayscale picture. To set the threshold picture, you may use the formula g(x,y)

$$g(x, y) = \begin{cases} 1 \;\; if\; f(x, y) > T \\ 0 \;\; if\; f(x, y) \leq \;\; T \end{cases} \text{ ----------- (9)}$$

3.5 Classification using improved VGG19

A number of medical imaging-specific improvements are part of the upgraded VGG19 network for lung cancer prediction. The network is fine-tuned using pictures relevant to lung cancer after being pre-trained on a big dataset like ImageNet. This allows it to better capture disease-related properties. The input layer is well-tuned to handle CT scans of varying sizes and types, preserving all the fine detail. Modified activation functions and additional filters in enhanced feature extraction layers allow for the detection of fine distinctions between normal and malignant tissues. Overfitting may be prevented and generalisation can be improved with the use of advanced regularisation methods including data augmentation, batch normalisation, and dropout. The last softmax layer generates a probability distribution, which gives confidence ratings for predictions, and custom fully connected layers are used to capture complicated patterns. All of these changes make the VGG19 network a better lung cancer classifier, which means it can be used more effectively for early detection and diagnosis.

The six primary building blocks of a VGG CNN are full-connected and multiple-connected convolutional layers. A 224*224*3 input is used with a 3*3 convolutional kernel. Typically, 16–19 layers are focused. Figure 1 shows the structure of the VGG-19 model.

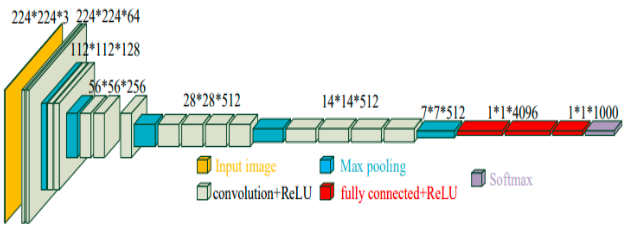

Figure 2: VGG-19 network model

There has been an improvement in network depth compared to typical convolutional neural networks. It outperforms using a single convolutional layer by alternating between many convolutional and non-linear activation layers. In order to improve feature extraction, downsampling, and activation, the layer structure employs Maxpooling and modifies the linear unit (ReLU). This means that the greatest value in the picture region is used as the pooled value of the area.

$$x_{p_j}^{(n)} = f\left(\tau_j^n\ down\left(x_j^{(n-1)}\right) + b_j^{(n)}\right) \text{ ------------ } (10)$$

IV. RESULTS AND DISCUSSION

We provide the findings and analysis of the integrated framework for lung cancer prediction here. The results provide information on the efficacy of each method, from improved vgg19 classification to noise reduction. In this section, we'll examine what these results mean and how the suggested framework may be improved upon, as well as its strengths and weaknesses.

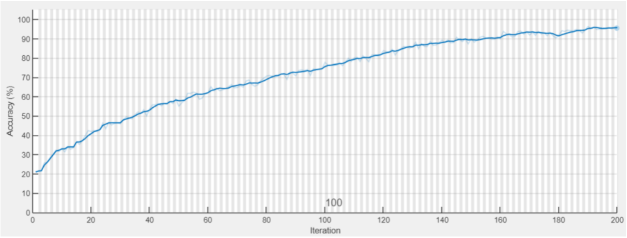

Figure 3: Training accuracy

Training accuracy is seen in figure 3. On one side, we can see the value of iterations, and on the other, we can see the training accuracy.

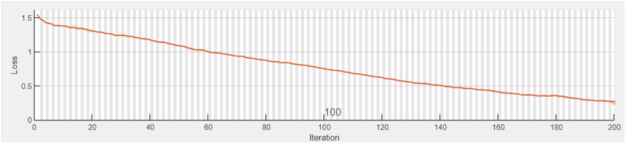

Figure 4: Training loss

Figure 4 displays the loss in training. Values for iterations and training losses are shown on the x and y axes, respectively.

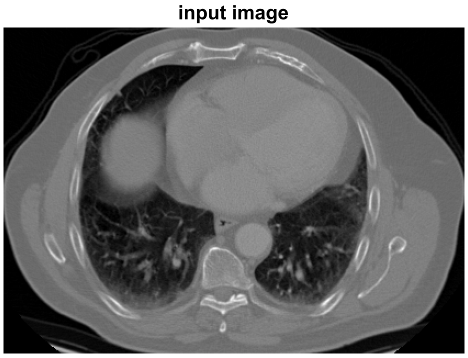

Figure 5: Input image

The picture that the lung cancer prediction algorithm uses as input is shown in Figure 5. The integrated system is built using this photograph as its starting data point. Potentially shown in the picture are lung tissues, which may vary in density, structure, texture, and other attributes.

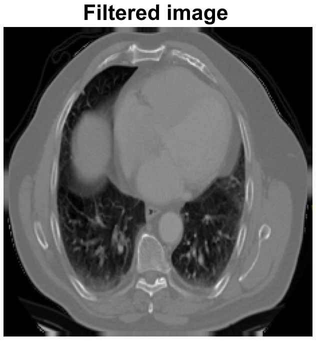

Figure 6: Filtered image

Figure 6 shows the filtered picture, which is an important part of the framework for predicting lung cancer. After applying a fuzzy-based median filter to the input picture shown in Figure 5, this image shows the outcome of the noise reduction process.

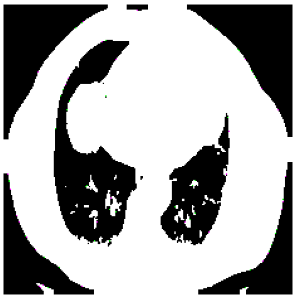

Figure 7: Threshold image

Figure 7, the Threshold picture, shows a crucial phase in the framework for predicting lung cancer. Following filtering and noise reduction, this image is the result of using a threshold-based multilate segmentation algorithm. Setting intensity thresholds as part of the thresholding procedure divides the picture into separate parts, allowing for the efficient isolation of possible malignant spots.

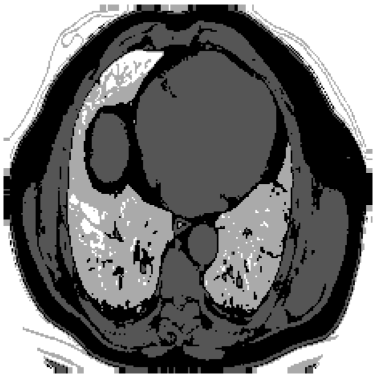

Figure 8: Segmented image

The Segmented picture, shown in Figure 8, is a major accomplishment in the framework for lung cancer forecasting. As shown in Figure 7, the filtered picture was subjected to the threshold-based multilate segmentation approach.

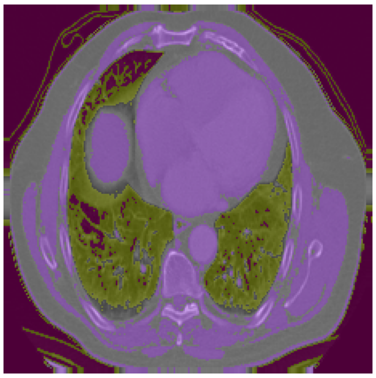

Figure 9: Predicted image stage 1

Figure 9 depicts the anticipated picture from Stage 1 of the framework for predicting lung cancer.

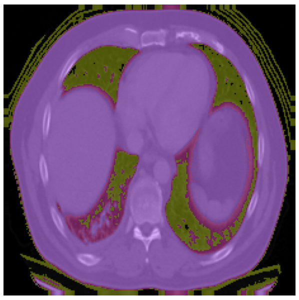

Figure 10: Predicted image stage 2

In a critical juncture in the lung cancer prediction paradigm, Figure 10 displays the anticipated image in stage 2. You can see the evolution of the predictions produced in stage 1 as they have been refined and improved upon throughout the many stages of the predictive model here.

Table 1: Threshold value comparison table

Algorithm	Threshold
DNN	0.5
RBF SVM	0.33
Adaboost	0.5
MLP	0.4
Improved vgg19	0.3

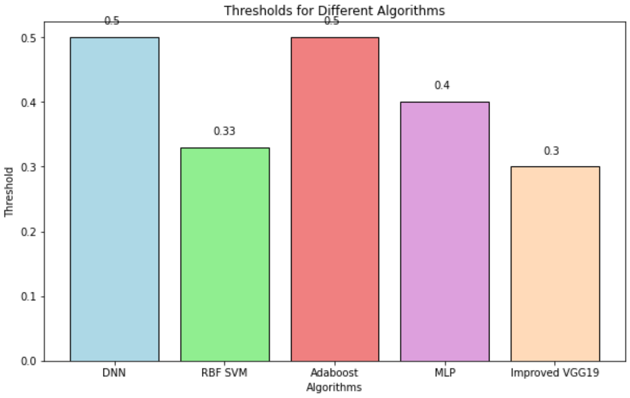

Figure 11: threshold value comparison chart

The thresholds for different algorithms used in lung cancer prediction vary, reflecting their sensitivity and decision criteria. The Deep Neural Network (DNN) and Adaboost both have a threshold of 0.5, indicating a balanced approach to classification where a prediction is made when the probability of the positive class reaches 50%. The RBF SVM has a lower threshold of 0.33, making it more sensitive to positive class predictions by requiring only a 33% probability. The MLP uses a threshold of 0.4, indicating it requires a 40% probability to classify a positive case, striking a balance between sensitivity and specificity. The Improved VGG19 network has the lowest threshold at 0.3, suggesting a higher sensitivity to positive predictions, as it classifies a positive case with just a 30% probability. These varying thresholds highlight the trade-offs between sensitivity and specificity across different machine learning models.

Table 2: Classification Performance metrics comparison table

	Algorithm	Sensitivity	Specificity	Accuracy
Existing authors	Banerjee, N., & Das, S.	91.32	91.20	92.00
	Baskar, S. et al.	90.31	92.31	91.11
Existing methods	DNN	80.52	75.32	83.47
	RBF SVM	82.61	81.91	83.22
	Adaboost	86.17	90.31	81.24
	MLP	87.77	84.18	82.53
Proposed method	Improved vgg19	98.31	94.07	97.9

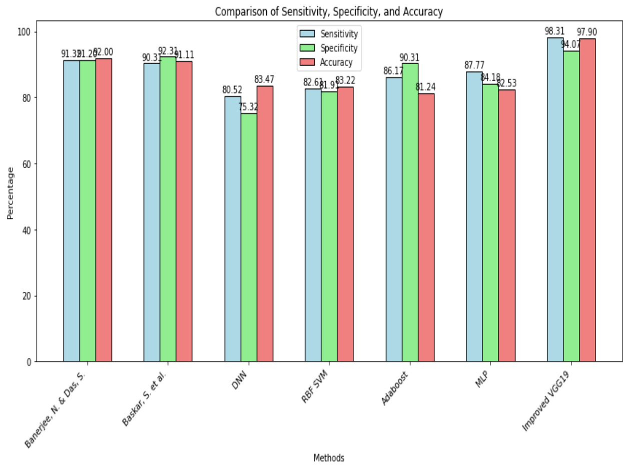

Figure 12: Performance metrics comparison chart

There are large disparities in the sensitivity, specificity, and accuracy of the several algorithms used to forecast lung cancer. Baskar et al. obtained a sensitivity of 90.31%, specificity of 92.31%, and accuracy of 91.11%; in contrast, Banerjee, N., & Das, S. attained a sensitivity of 91.32%, specificity of 91.20%, and accuracy of 92.00%. When looking at current methods, we can see that DNN scored 80.52% for sensitivity, 75.32% for specificity, and 83.47% for accuracy. RBF SVM scored 82.61% for sensitivity, 81.91% for specificity, and 83.22% for accuracy. Adaboost scored 86.17% for sensitivity, 90.31% for specificity, and 81.24% for accuracy. Finally, MLP achieved 87.57% for sensitivity, 84.18% for specificity, and 82.53% for accuracy. Using Improved VGG19, the suggested technique achieved the best results compared to others, with a sensitivity of 98.31%, specificity of 94.07%, and amazing accuracy of 97.9%. This clearly indicates its great potential for reliably and precisely predicting instances of lung cancer.

V. CONCLUSION

In conclusion, this study presents a comprehensive and highly effective methodology for lung cancer prediction that integrates advanced image processing techniques with an improved VGG19 network. The use of a fuzzy-based median filter for noise removal, CLBP for feature extraction, and threshold-based Mutilate Segmentation for precise region isolation significantly enhances the quality and reliability of the input data. The optimized VGG19 network, tailored specifically for lung cancer detection, achieves an outstanding accuracy of 97.9%, showcasing its superior capability in recognizing disease-specific patterns and reducing false-positive rates. Experimental results validate the efficacy of the proposed framework, highlighting its potential for clinical applications and emphasizing the critical role of integrating sophisticated image processing methods with.

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