Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 11, No.1, November : 2024

LUNG CANCER DETECTION USING MACHINE LEARNING

Dr.Naresh Thoutam Dept. of Computer Engineering,Sandip Institute of Technology & Research Center,Nashik, India naresh.thoutam@sitrc.org
Prof.P.G.Patil Dept. of Computer Engineering,Sandip Institute of Technology & Research Center,Nashik, India pramod.patil@sitrc.org
Vipul Lokhande Assistant Professor, Dept. of Computer Engineering Sandip Institute of Technology & Research Center,Nashik, India lokhandevipul245@gmail.com
Aaditi Baviskar Dept. of Computer Engineering,Sandip Institute of Technology & Research Center,Nashik, India aditibaviskar26@gmail.com
Vidur Diwate Dept. of Computer Engineering,Sandip Institute of Technology & Research Center,Nashik, India vidurdiwate@gmail.com
Anmol S Budhewar Dept. of Computer Engineering,Sandip Institute of Technology & Research Center,Nashik, India anmolsbudhewar@gmail.com

ABSTRACT
The Main Objective of this research paper is to find out the early stage of lung cancer and explore the accuracy levels of various machine learning algorithms. After a systematic literature study, we found out that some classifiers have low accuracy and some arehigher accuracy but difficult to reached nearer of 100%. Low accuracy and high implementation cost due to improper dealing with DICOM images. For medical image processing many different types of images are used but Computer Tomography (CT) scans are generally preferred because of less noise. Deep learning is proven to be the best method for medical image processing, lung nodule detection and classification, feature extraction and lung cancer stage prediction. In the first stage of this system used image processing techniques to extract lung regions. The segmentation is done using K Means. The features are extracted from the segmented images and the classification are done using various machine learning algorithm. The performances of the proposed approaches are evaluated based on their accuracy, sensitivity, specificity and classification time.

Keywords:
Structural Co-occurrence Matrix (SCM), Classifer, Data Set, ROC curve, Malignant nodule, Benign nodule .

I. INTRODUCTION
The cause of lung cancer is still unknown and has become impossible to prevent; therefore, the only way to treat lung cancer is early detection. The size of the tumor and the rate at which it spreads dete rmine the stage of the cancer.[1] Cancer is common all over the world.[2] Mortality and health probl ems are common in many countries, and the 5year survival rate is only 10% to 16% . Sometimes the nodules are not obvious and it takes a trained eye and a lot of time to make a diagnosis.[3]In addition , most lung nodules are not cancerous because theycan be caused by noncancerous growths, scars, or diseases [4]. machine learning-based lesion identification, two primary techniques are detection and segmentation. Detection focuses on classifying lesions at a broader, neighborhood level, while segmentation works at the pixel level, providing more detailed information.[5] Segmentation is often more useful in clinical settings, as it allows for precise mapping of lesions at the pixel scale, improving diagnostic accuracy. By identifying lesions at this granular level, clinicians can more easily monitor changes in size and shape over time, and use the lesion's shape as a reference for more accurate detection and assessment of disease progression. Although many researchers use machine learning to study. [6]The problem with this method is that many manual measurements are required to evaluat e the performance, making it difficult to produce better results [7]. Classification is an important part of the computation that groups images according to their similarity [8] [9]. In the cancer cell example

, most of the cells overlap each other. so catch it earlyCancer is a more challenging task [10] [11]. Af ter extensive research, we found that the combination of components performs well compared to othe r machine learning algorithms [12]. Existing CAD methods for early cancer detection with the help o f CT images are not satisfactory due to low sensitivity and negative precision (FPR).

II.PROPOSED WORK

In our approach, we begin by pre-processing the data to extract and select the most relevant features, discarding the less important ones. After feature selection, the model is trained and tested using several algorithms, including CNN, Random Forest, and a hybrid neural network model. We then evaluate and compare the performance of these different models using a variety of performance metrics, allowing us to assess the efficiency and accuracy of each algorithm in the given context.

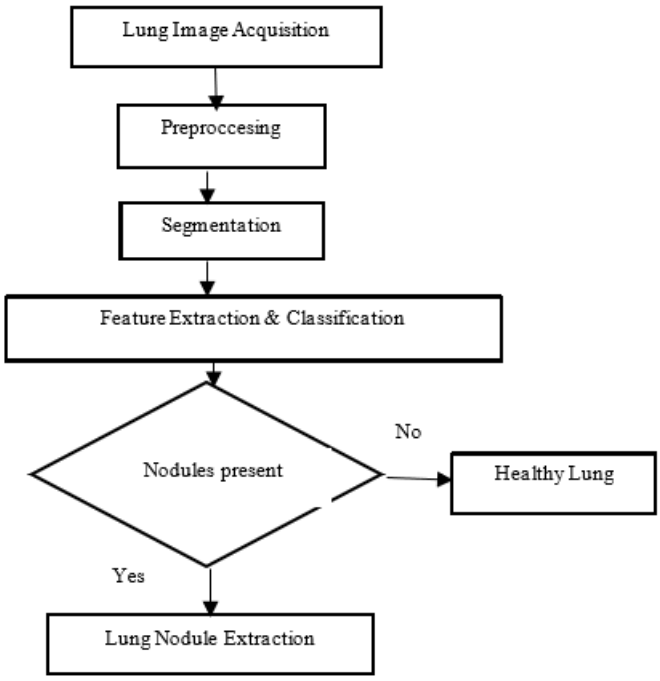

Figure 1. PROPOSED SYSTEM

❖ Dataset/Data Collection

Dataset Provide dataset (This means that the data collected should be made uniform and understandable for a machine that doesn't see data the same way as humans do.)

❖ Data pre-processing

Pre-processing is the initial step in lung cancer detection, where the CT scan image is taken as input and prepared for analysis. The primary task in this phase is image de-noising, which involves applying techniques to remove noise or artifacts that may obscure critical details in the image. By eliminating these distortions, de-noising enhances the clarity of lung structures, making it easier to detect abnormalities such as nodules. The result is a higher-quality, noise-free image that significantly improves the accuracy and reliability of subsequent processes like segmentation and feature extraction, making it a crucial stage in the detection of lung cancer.

❖ Data segmentation

region-based segmentation is applied to divide the CT scan image into similar and dissimilar regions. This process helps isolate areas of interest, such as potential lung abnormalities. Otsu's segmentation technique is used for thresholding, which simplifies the image by converting it from grayscale to a binary (black and white) format. This conversion reduces the image's complexity by representing it in just two values, making storage and processing faster and more efficient compared

to working with a grayscale image that has 256 intensity levels. The resulting segmented image retains key features while facilitating easier analysis in the lung cancer detection process.

❖ Feature Extraction

feature extraction is carried out using the GLCM (Gray Level Co-occurrence Matrix) algorithm, which focuses on extracting the textural characteristics of the CT scan image. The GLCM analyzes the spatial relationships between pixel intensities in the image, allowing it to capture important textural details such as contrast, correlation, energy, and homogeneity. These features provide valuable information about the patterns and structures within the lung tissues, helping to differentiate between normal and abnormal regions. By extracting these key textural features, the GLCM algorithm plays a crucial role in enhancing the accuracy of lung cancer detection

❖ Classification

Convolutional Neural Networks (CNNs) play a crucial role in image classification, especially in medical imaging for tasks like identifying and localizing cancerous regions in CT scans. CNNs work by distinguishing different data classes, aiming to create a clear separation between them using an optimal hyperplane. By maximizing this margin, CNNs ensure that no data points from different classes overlap, thereby enhancing classification accuracy. This method is particularly important for the precise detection and localization of cancerous versus non-cancerous areas, aiding in effective lung cancer diagnosis.

III.MODEL ARCHITECTURE

We utilized a CT scan image dataset from Kaggle, consisting of DICOM format slices. The process began with exploratory data analysis (EDA) to evaluate the dataset's quality and consistency. Through various graphs and visualizations, we gained a better understanding of the image distribution and characteristics. In the data pre-processing phase, we segmented the images, generated 3D visualizations, and rendered scan volumes to analyze the internal structures. From these images, we extracted key attributes like imaging modality, dimensions, pixel spacing, slice position, and other relevant metadata provided by the DICOM format, which made feature extraction more efficient. After feature extraction, we divided the dataset into training and testing sets, reserving 20% for testing. To understand the architecture, we generated a model summary to analyze the layers and parameters. The performance of the model was visualized using Matplotlib, plotting accuracy and loss during training and validation. After training, we examined the important characteristics extracted by the model to evaluate its effectiveness in classifying the CT scan images.

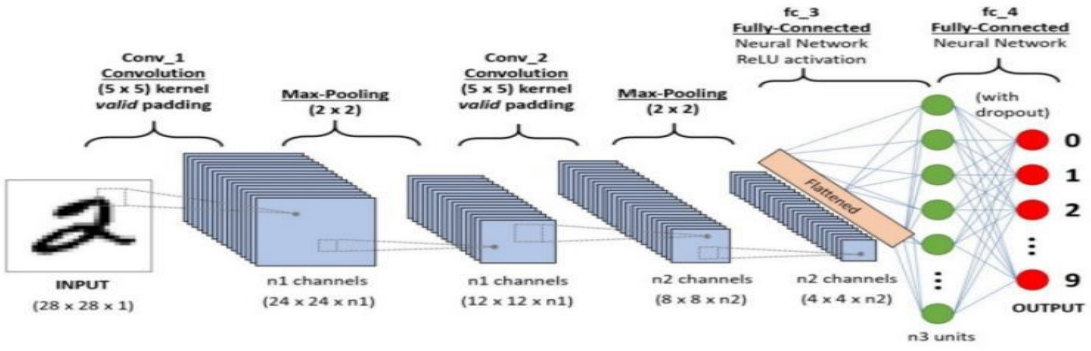

Figure 2 .CNN ARCHITECTURE

CNN: CNNs consist of convolutional layers, which contain trainable parameters, followed by pooling layers. These layers, particularly max pooling and average (or optimized) pooling, are critical for reducing dimensionality while retaining essential features.

IV. LITERATURE

1. In the paper by Pankaj Nanglia and Sumit Kumar et al. [1], a novel approach called the "kernel feature selection classifier" was proposed, which integrates Support Vector Machine (SVM) with a

feedforward backpropagation neural network (FFBPNN) to improve distribution rates. Their methodology involved three key stages. The first stage focused on data distribution, while the second stage used SURF technology for feature extraction, followed by genetic algorithm optimization. The third stage involved classification using FFBPNN. Additionally, a sensitivity analysis using multicenter data was conducted by Dang et al., who focused on two groups: diameter and pathological findings. The diameter was divided into three categories: 0-10mm, 10-20mm, and 20-30mm. In the 0-10mm group, sensitivity was 85.7% (95% CI, 70.8% to 100%) and specificity was 91.1% (95% CI, 86.8% to 95.2%). For the 10-20mm group, sensitivity was 85.7% (95% CI, 77.1% to 94.3%) and specificity was 90.1% (95% CI, 84.8% to 95.4%). The algorithm achieved the highest accuracy of 85.7% for adenocarcinoma and 65.0% for squamous cell carcinoma.

2. Sangam Borkar's research focused on distinguishing between normal and abnormal lung images by addressing noise. Using mathematical morphology functions, lung segmentation was enhanced to detect tumor regions. Three geometric features—area, perimeter, and eccentricity—were extracted from the segmented regions and classified using an SVM classifier. K. Rajeswari explored pulmonary nodule detection and malignancy prediction using lung CT images, leveraging datasets such as LIDC_IDRI, LUNA16, and Data Science Bowl 2017. Their experiments, conducted on a CUDA-enabled GPU Tesla K20, employed a U-Net architecture for segmentation and a 3D multi-image VGG architecture for nodule classification and malignancy prediction, achieving an accuracy of 95.66%, a loss of 0.09, a dice coefficient of 90%, and a prediction loss of 38%.

3. In their study, Moffy Vas and Amita Dessai [2] concentrated on classifying cancerous and non-cancerous images. Their pre-processing involved removing irrelevant parts of lung CT scans and using a median filter to eliminate salt-and-pepper noise. Mathematical morphology was applied to refine lung segmentation and detect tumor areas. They extracted seven features—strength, correlation, diversity, homogeneity, difference entropy, correlation, and contrast—which were fed into a feedforward neural network using a backpropagation algorithm for classification. The algorithm minimized the error function using a gradient descent method, achieving a training accuracy of 96%, a testing accuracy of 92%, a sensitivity of 88.7%, and a specificity of 97.1%.

4. In paper [3], Radhika P R and Rakhi A S Nair focused on predicting and classifying medical imaging data using datasets from the UCI Machine Learning Repository and data.world. They compared various machine learning algorithms, concluding that SVM provided the highest accuracy at 99.2%, followed by Decision Tree at 90%, Naive Bayes at 87.87%, and Logistic Regression at 66.7

5. In paper Vaishnavi D., Arya K. S., Devi Abirami T., and M. N. Kavitha [4], a lung cancer detection algorithm was developed. During the pre-processing stage, they applied the Dual-tree Complex Wavelet Transform (DTCWT), where the wavelet is discretely sampled. For texture analysis, they utilized the Gray Level Co-occurrence Matrix (GLCM), a second-order statistical method that tabulates how different combinations of gray levels occur together in an image, measuring intensity variation at specific pixels. They employed a Probability Neural Network (PNN) classifier, which was evaluated based on training performance and classification accuracy, offering quick and precise classification results.

6. In paper K. Mohanambal and Y. Nirosha [5], they used a structural co-occurrence matrix (SCM) to extract features from images and classified them as malignant or benign based on these features. The Support Vector Machine (SVM) classifier was applied to categorize lung nodules according to their malignancy levels, ranging from 1 to 5.

V.CONCLUSION

Lung cancer CAD includes steps such as preprocessing, nodule detection, nodule segmentation, featu re extraction, and classification of benign and malignant nodules. Once a nodule is detected and seg mented, the feature extraction process begins. Feature extraction techniques are used to extract the ne cessary features for classification from the segmented nodules. Based on the extracted results, the pro duct is used to classify nodules as benign or malignant. Future research should focus on addressing

ethical considerations, ensuring data privacy, and validating these models in real-world clinical settings to maximize their impact on healthcare. The goal of this study is to predict lung cancer detection based on image. The project is set up in such a way that the system takes the image as input and outputs a prediction of lung cancer detection.

REFERENCES

[1] N. Camarlinghi, "Automatic detection of lungnodules in computed tomography images: Training and validation of algorithms using public research
[2] R. L. Siegel, K. D. Miller, and A. Jemal,
"Cancer statistics, 2016", CA, Cancer J. Clin.,vol. 66, no. 1, pp. 730, 2016.
[3] Detecting and classifying nodules in Lung CT scans, **http://modelheelephant.blogspot.com/2017/11/detecting-and-classifying-nodules- in.html,2017**.
[4] Diego Riquelme and Moulay A. Akhloufi ,"DeepLearning for Lung Cancer Nodules Detection andClassification in CT Scans",www.mdpi.com,2020.
[5] Anita Chaudhary, Sonit Sukhraj Singh,"Lung Cancer Detection on CT Images by using Image Processing",IEEE,2012.
[6] Gawade Prathamesh Pratap, R.P. Chauhan,
"Detection of Lung Cancer Cells using Image Processing Techniques", International Conference on Power Electronics, Intelligent Control and Energy Systems(ICPEICES),2016.
[7] Pooja R. Katre,Dr. Anuradha Thakare, "Detection of Lung Cancer Stages using Image Processing and Data Classification Techniques", International Conference for Convergence in Technology,IEEE,2017
[8] Rituparna Sarma, Yogesh Kumar Gupta"Acomparative study of new and existing segmentation techniques",ICCRDA,2020.
[9] Eali Stephen Neal Joshua1*, Midhun Chakkravarthy1, Debnath Bhattacharyya2 ," An Extensive Review on Lung Cancer Detection Using Machine Learning Techniques: A Systematic Study",International Information and Engineering Technology Association (IIETA),2020.
[10] Pankaj Nanglia,Sumit Kumar, Aparna N. Mahajan, Paramjit Singh, Davinder Rathee,"Ahybrid algorithm for lung cancer classificationusing SVM and Neural Networks", The Korean Institute of Communication and Information Science(KICS),2020.Also available at www.elsevier.com/locate/icte.
[11] Chao Zhang,Xing Sun, Kang Dang et all "Towardan Expert Level of Lung Cancer Detection and Classification
[12] Kuru villa, Jinsa, and K. Gunavathi. "Lung Cancer Classification Using Neural Networks for CT Images. " Computer Methods and Program s in Biomedicine 113.1 (2014)
[13] Wang, Xing, et al. "an Appraisal of Lung Nodules Automatic Classification Algorithm s for CT Images." Sensors (Basel)
[14] Gomathi M. Thangaraj P: Computer aided medical diagnosis system for detection of lung cancer nodules: a survey. Int J Comput Intell Res 2009.
[15] D. F. Sheehan et al., "Lung cancer costs by treatment strategy and phase of care among patients enrolled in Medicare," (in English), Cancer Med-Us, vol. 8, no. 1, pp. 94-103, Jan 2019, doi: 10.1002/cam4.1896