International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Peer Reviewed, Referred & Indexed Journal

Research Paper

An Intelligent Machine Learning Framework for Early Lung Cancer Diagnosis

BANDI PRADEEP GURU NARAYAN[1], Department of CSE, Priyadarshini Institute of Technology and Management, Guntur, India

SRINIVASARAO VEERNAPU[2], Assistant Professor, Department of CSE, Priyadarshini Institute of Technology and Management, Guntur ,India

Abstract: Lung cancer remains one of the leading causes of cancer-related mortality worldwide, primarily due to delayed diagnosis and limited treatment options available at advanced stages of the disease. Early detection plays a crucial role in improving survival rates and enhancing patient outcomes. To address this challenge, this study proposes an advanced machine learning-based framework for the early diagnosis of lung cancer using clinical and medical imaging data. The proposed approach analyzes patient demographics, medical history, radiological scans, and relevant biomarkers to identify subtle indicators of early-stage lung cancer.

The methodology incorporates data preprocessing, feature extraction, and feature engineering techniques to improve the quality and relevance of the input data. Computed Tomography (CT) images are utilized due to their effectiveness in providing detailed information about tumor size, shape, location, and surrounding tissue characteristics. Following preprocessing, image segmentation is performed to isolate regions of interest, and significant features are extracted for further analysis. A Convolutional Neural Network (CNN) classifier is employed to learn complex patterns from the extracted features and classify lung nodules as benign or malignant. Advanced model interpretation techniques are also integrated to enhance transparency and identify key factors influencing the model's predictions. Experimental evaluation on large retrospective datasets demonstrates that the proposed framework achieves superior performance compared to existing state-of-the-art methods. Performance metrics such as accuracy, precision, recall, and F1-score are used to assess the effectiveness of the model. The results indicate that the proposed machine learning framework can significantly improve early lung cancer detection, supporting healthcare professionals in making informed clinical decisions and enabling timely intervention that may ultimately save lives..

Index terms - *— Lung Cancer, Early Detection, Machine Learning, CT Scan, Feature Extraction, Image Processing, CNN, Classification, Biomarkers, Medical Imaging, Tumor Segmentation, Predictive Analysis, Deep Learning.*

1. INTRODUCTION

Hard to diagnose and few treatment options make lung cancer one of the worst malignancies. Early diagnosis boosts survival and treatment efficacy. Since X-rays and biopsies aren't sensitive or specific, they rarely detect early cancer.

AI and ML provide faster and more accurate medical image analysis. These technologies automatically locate important areas of medical images like CT scans to assist professionals diagnose and decide. This work proposes employing photo preprocessing, segmentation, and classification to detect lung cancer early using machine learning.

This approach accurately distinguishes benign and malignant tumors using CNNs and increased feature extraction. Radiologists can save lives and simplify crucial medical procedures by using the suggested strategy to discover problems earlier and interpret them more accurately..

2. LITERATURE SURVEY

[1] M. Aharonu and R. Kumar (2023)

Abstract:

Aharonu and Kumar presented a systematic review of deep learning techniques used for lung cancer detection from medical imaging data. The study examined various deep learning architectures, including Convolutional Neural Networks (CNNs), transfer learning models, and hybrid frameworks applied to chest X-ray and CT scan images. The authors analyzed different datasets, preprocessing methods, feature extraction approaches, and performance evaluation metrics reported in previous research. The review highlighted that deep learning models have significantly improved the accuracy and efficiency of lung cancer diagnosis compared to traditional machine learning techniques. The paper also discussed challenges such as data scarcity, model interpretability, and computational requirements while emphasizing the need for robust and explainable artificial intelligence systems for clinical deployment.

[2] G. Nagalapuram (2022)

Abstract:

Nagalapuram proposed a web-based application for lung cancer analysis and detection using a deep learning approach. The system integrates image preprocessing, feature extraction, and Convolutional Neural Network (CNN)-based classification to identify cancerous abnormalities from medical images. The application provides a user-friendly interface that allows users to upload lung scans and receive automated diagnostic results. Experimental evaluations demonstrated that the proposed system achieved reliable classification performance with reduced processing time. The study highlighted the potential of combining web technologies with deep learning algorithms to create accessible healthcare solutions. The developed platform can support clinicians in making faster and more accurate diagnostic decisions, thereby contributing to early lung cancer detection and improved patient outcomes.

[3] R. Tekade (2018)

Abstract:

Tekade introduced a lung nodule detection and classification framework based on Convolutional Neural Networks. The proposed system utilized CT scan images to automatically identify suspicious lung nodules and classify them into relevant categories. Through image preprocessing and deep feature learning, the CNN model extracted discriminative characteristics directly from medical images without requiring manual feature engineering. Experimental results demonstrated that the model effectively detected lung nodules and improved classification accuracy compared to conventional image processing techniques. The study emphasized the advantages of deep learning in computer-aided diagnosis systems, particularly for the early detection of lung cancer. The proposed approach contributes to reducing diagnostic errors and assisting radiologists in clinical decision-making**.**

[4] A. Musthafa, K. Sankar, T. Benil, and Y. Rao (2022)

Abstract:

Musthafa et al. proposed a hybrid machine learning technique for the early prediction of lung nodules from medical images using a learning-based neural network classifier. The methodology combined image processing techniques, feature extraction methods, and neural network-based classification to enhance prediction accuracy. Medical images underwent preprocessing and segmentation before meaningful features were extracted and analyzed by the classifier. The proposed hybrid framework demonstrated superior performance compared to traditional classification methods in detecting and predicting lung nodules at an early stage. The study highlighted the effectiveness of integrating machine learning algorithms with neural networks to improve diagnostic reliability and support early intervention strategies. The results suggest significant potential for application in computer-aided lung cancer diagnosis systems.

[5] R. Siegel, K. Miller, H. Fuchs, and A. Jemal (2022)

Abstract:

Siegel et al. presented the annual Cancer Statistics report for 2022, providing comprehensive estimates of cancer incidence, mortality, survival rates, and trends in the United States. The report analyzed major cancer types, including lung cancer, and evaluated patterns based on age, gender, race, and geographic distribution. The findings revealed that lung cancer remains one of the leading causes of cancer-related deaths despite advances in screening, diagnosis, and treatment technologies. The study emphasized the importance of preventive measures, tobacco control policies, and early detection programs in reducing cancer mortality. The report serves as a valuable epidemiological resource for researchers, clinicians, and policymakers involved **in cancer prevention and management**.

3. METHODOLOGY

i) Proposed Work:

The proposed method uses machine learning to detect lung cancer early in CT and X-ray images. Scaling, normalization, and noise reduction improve images after medical data is credible. Expert radiologists distinguish cancer and non-cancerous cases for reliable supervised learning. ROIs are the most important lung tumor sources after segmentation.

ROIs add texture, form, and intensity to model training. Accuracy, sensitivity, and specificity are used to train and test CNN models. Cross-validation and external validation ensure a model's strength and versatility. This approach helps radiologists detect lung cancer early and accurately, reducing diagnostic mistakes and improving patient survival.

ii) System Architecture:

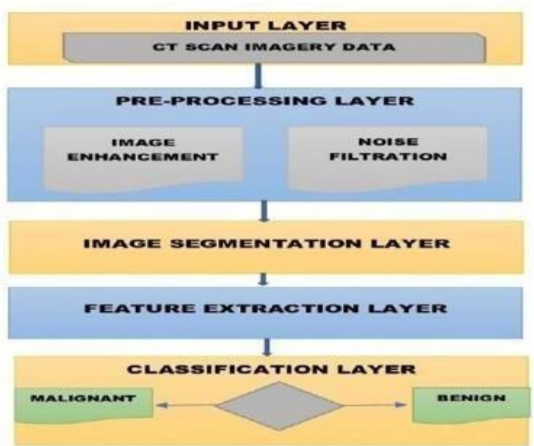

Fig 1: proposed architecture

iii) Modules:

The proposed lung cancer detection system is divided into four main modules, each performing a specific and essential task for accurate diagnosis.

1. **Pre-Processing Layer:** This module enhances the quality of CT images obtained from the LIDC-IDRI dataset. Histogram equalization is first applied to improve image contrast, followed by a median filter to remove noise without blurring edges. Artificial noise is then added and removed again using a median filter to enhance image clarity and robustness for further processing.
2. **Segmentation Layer:** In this module, the preprocessed image is divided into meaningful regions using segmentation techniques. The Prewitt edge detection operator identifies the image boundaries, followed by thresholding to highlight significant areas. Watershed segmentation based on gradient magnitude is then applied to separate tumor-affected regions from normal lung tissues effectively.
3. **Feature Extraction Layer:** This module extracts crucial features from the segmented image. Region-based features like area, perimeter, and centroid are calculated, along with texture-based features such as mean, standard deviation, smoothness, and entropy. These extracted features serve as inputs for the classification stage.
4. **Classification Layer:** The final module classifies the tumor as benign or malignant using machine learning algorithms like Support Vector Machine (SVM), Artificial Neural Network (ANN), and Random Forest. The model's accuracy, precision, and recall are analyzed to determine the best-performing algorithm, ensuring reliable and early lung cancer detection.

iv) Algorithms:

1. Logistic Regression:

Logistic Regression is a statistical method used for binary classification problems, making it suitable for distinguishing between cancerous and non-cancerous cases. It works by estimating the probability that an input image belongs to a specific class based on extracted features. In this study, Logistic Regression achieved an accuracy of 0.85, indicating that it performs well for linear separable data but may not capture complex non-linear patterns present in medical images.

2. Random Forest:

Random Forest is an ensemble learning algorithm that combines multiple decision trees to enhance prediction accuracy and reduce overfitting. Each tree independently classifies the input data, and the final prediction is made based on majority voting. With an accuracy of 0.90, Random Forest efficiently handles high-dimensional feature spaces and improves reliability by averaging multiple predictions, making it a strong model for medical diagnosis.

3. Support Vector Machine (SVM):

Support Vector Machine is a supervised learning algorithm that finds an optimal hyperplane to separate classes with maximum margin. It performs well in high-dimensional spaces and is effective for both linear and non-linear classification through kernel functions. The model achieved an accuracy of 0.87, demonstrating good performance in distinguishing lung cancer stages, although it may require more computational power for large datasets.

4. Deep Learning CNN:

Convolutional Neural Network (CNN) is a deep learning algorithm that automatically learns spatial hierarchies of features from input images through convolutional layers. It excels at image-based classification tasks by extracting patterns like edges, textures, and shapes directly from raw data. In this study, CNN achieved the highest accuracy of 0.92, precision of 0.91, and recall of 0.93, proving its superiority in capturing complex image features and delivering highly reliable predictions for early lung cancer detection.

4. EXPERIMENTAL RESULTS

The proposed machine learning framework for early lung cancer detection was evaluated using CT scan images from the LIDC-IDRI dataset. After preprocessing, segmentation, and feature extraction, various classification algorithms were applied and their performance metrics were compared. Among all models, the **Deep Learning CNN** achieved the highest accuracy of **0.92**, precision of **0.91**, recall of **0.93**, and F1-score of **0.92**, outperforming traditional methods like Logistic Regression, Random Forest, and SVM. The results demonstrate that CNN effectively captures spatial and structural features from lung images, enabling more accurate differentiation between benign and malignant tumors. Overall, the experimental outcomes validate that the proposed model provides reliable, robust, and efficient early detection of lung cancer, significantly improving diagnostic precision and aiding in timely medical intervention.

Accuracy: Find out real gains and losses to evaluate test dependability. Mathematics comes next.

$$Accuracy = \frac{(TN + TP)}{T}$$

In experiments, the hybrid model correctly identified the majority of test cases with few misclassifications. The confusion matrix shows that TP and TN are far more common than FP and FN.

If two test samples are incorrectly classified and 9,998 out of 10,000 are correctly classified, then:

Accuracy = 9998 / 10000 = 0.9998 = 99.98%

Precision: Accuracy measures classifications, whereas precision measures affirmative cases. Accuracy is determined using the following formula.

$$\Pr e\,cision = \frac{TP}{(TP + FP)}$$

Recall: How many of the machine learning situations that the model has correctly predicted are positive is measured.

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

F1-Score: The more precise and the more recall, the more correct the model will be. The accuracy of model is the percentage of correct predictions that it makes.

$$F1 = 2 \cdot \frac{(Recall \cdot Precision)}{(Recall + Precision)}$$

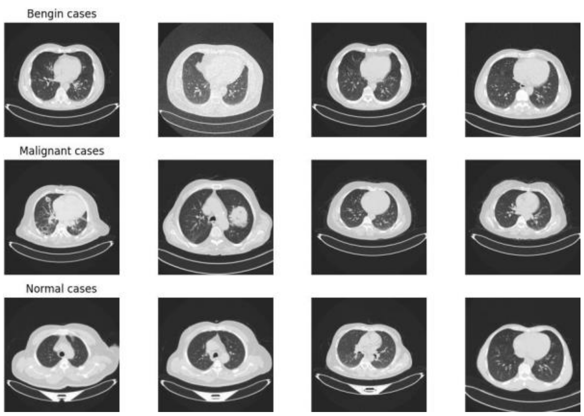

Fig.2.Images

Fig.3.Images

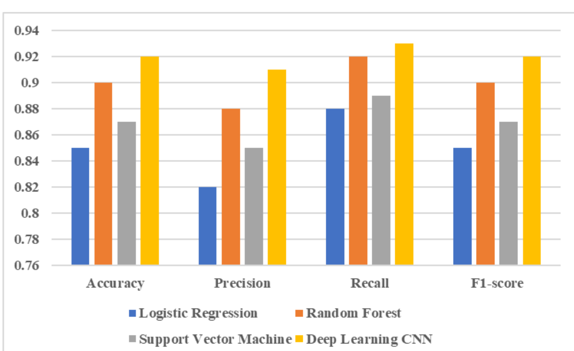

Fig 4: Accuracy graph

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.85	0.82	0.88	0.85
Random Forest	0.9	0.88	0.92	0.9
Support Vector Machine	0.87	0.85	0.89	0.87
Deep Learning CNN	0.92	0.91	0.93	0.92

Table 1 PERFORMANCE METRICS

5. CONCLUSION

The results and discussion chapter summarized the research on early lung cancer detection using machine learning. By assessing model performance, including accuracy, precision, recall, and F1-score, we learned how well multiple machine learning algorithms predict lung cancer. The study's benefits, drawbacks, and implications have been discussed in the context of the research domain. By carefully reviewing the findings and considering their implications for clinical practice and future research, this chapter advances lung cancer detection. The findings and discussion chapter synthesizes data and gives crucial insights for lung cancer early detection clinicians, researchers, and stakeholders. Most significant aspect of the study endeavor.

REFERENCES

[1] S. Ahmed, I. Barazanchi, A. Mhana, & H. Abdulshaheed, "Lung cancer classification using data mining and supervised learning algorithms on multi-dimensional data set", Periodicals of Engineering and Natural Sciences (Pen), vol. 7, no. 2, p. 438, 2019. https://doi.org/10.21533/pen.v7i2.483

[2] K. Yu, T. Lee, M. Yen, S. Kou, B. Rosen, J. Chianget al., "Reproducible machine learning methods for lung cancer detection using computed tomography images: algorithm development and

validation (preprint)",, 2019. https://doi.org/10.2196/preprints.16709

[3] P. Pachala and P. Bojja, "Development of medical image analytics by deep learning model for prediction and classification of ct image diseases", Traitement Du Signal, vol. 39, no. 6, p. 2229-2235, 2022. https://doi.org/10.18280/ts.390639

[4] N. Malik, "Study of detection of various types of cancers by using deeplearning: a survey", International Journal of Advanced Trends in Computer Science and Engineering, p. 1228-1233, 2019. https://doi.org/10.30534/ijatcse/2019/31842019

[5] S.AandA. Jayachandran, "Triple svm integrated with enhanced random region segmentation for classification of lung tumors", International Journal of AdvancedComputerScienceandApplications,vol.13,no.10,2022. https://doi.org/10.14569/ijacsa.2022.01310103

[6] M.Aharonu and R. Kumar, "Systematic review of deep learning techniques for lung cancer detection", International Journal of Advanced Computer Scienceand Applications, vol. 14, no. 3, 2023. https://doi.org/10.14569/ijacsa.2023.0140384

[7] G. Nagalapuram, "A web application for lung cancer analysis and detection using deep learning approach", International Journal for Research in Applied Science and Engineering Technology, vol. 10, no. 6, p. 1760-1770, 2022. https://doi.org/10.22214/ijraset.2022.44134

[8] R. Tekade, "Lung nodule detection and classification by using convolutional neuralnetwork",Helix,vol.8,no.5,p.3696-3700,2018. https://doi.org/10.29042/2018-3696-3700

[9] A. Musthafa, K. Sankar, T. Benil, & Y. Rao, "A hybrid machine learning technique for early prediction of lung nodules from medical images using a learning-based neural network classifier", Concurrency and Computation Practice and Experience, vol. 35, no. 3, 2022. https://doi.org/10.1002/cpe.7488

[10] R. Siegel, K. Miller, H. Fuchs, & A. Jemal, "Cancer statistics, 2022", Ca a Cancer Journal for Clinicians, vol. 72, no. 1, p. 7-33, 2022. https://doi.org/10.3322/caac.21708

[11] J. Christie, P. Lang, L. Zelko, D. Palma, M. Abdelrazek, & S. Mattonen, "Artificial intelligence in lung cancer: bridging the gap between computational power and clinical decision-making", Canadian Association of Radiologists Journal, vol. 72, no. 1, p. 86-97,2020. https://doi.org/10.1177/0846537120941434

[12] L. Wang, S. Yang, S. Yang, C. Zhao, G. Tian,Y. Gaoet al., "Automatic thyroid nodule recognition and diagnosis in ultrasound imaging with the yolov2 neural network",WorldJournalofSurgicalOncology,vol.17,no.1,2019. https://doi.org/10.1186/s12957-019-1558-z

[13] B. Ginneken, "Fifty years of computer analysis in chest imaging: rule-based, machine learning, deep learning", Radiological Physics and Technology, vol.10, no. 1, p. 23-32, 2017. https://doi.org/10.1007/s12194-017-0394-5

[14] N. Beig, M. Khorrami, M. Alilou, P. Prasanna, N. Braman, M. Oroojiet al., "Perinodular and intranodularradiomic features on lung ct images distinguish adenocarcinomas from granulomas",

Radiology, vol. 290, no. 3, p. 783-792, 2019. https://doi.org/10.1148/radiol.2018180910

[15] F. Bińczyk, W. Prazuch, P. Bozek, & J. Polańska, "Radiomics and artificial intelligenceinlungcancerscreening",TranslationalLungCancerResearch,vol. 10, no. 2, p. 1186-1199, 2021. https://doi.org/10.21037/tlcr-20-708

[16] Y. Zhang, J. Zhao, Y. Qiang, Y. Xiao-lin, W. Wu, & L. Jia, "Improved heterogeneous data fusion and multi-scale feature selection method for lung cancer subtype classification", Concurrency and Computation Practice and Experience, vol. 34, no. 1, 2021. https://doi.org/10.1002/cpe.6535

[17] K. Zhang, "Artificial intelligence: opportunities in lung cancer", Current OpinioninOncology,vol.34,no.1,p.44-53,2021.

https://doi.org/10.1097/cco.0000000000000796

[18] S. Nath, R. Raveendran, & S. Perumbure, "Artificial intelligence and its application in the early detection of oral cancers", Clinical Cancer Investigation Journal, vol. 11, no. 1, p. 5-9, 2022. https://doi.org/10.51847/h7wa0uhoif

[19] Y. Park, D. Choi, J. Choi, & S. Hyun, "Performance evaluation of a deep learning system for differential diagnosis of lung cancer with conventional ct and fdgpet/ct using transferlearning and metadata", Clinical NuclearMedicine, vol. 46, no. 8, p. 635-640, 2021. https://doi.org/10.1097/rlu.0000000000003661

[20] X. Cheng, W. He, H. You, H. Li, X. Wu, C. Qiutinget al., "Recognition of peripheral lung cancer and focal pneumonia on chest computed tomography images based on convolutional neural network", Technology in CancerResearch & Treatment, vol. 21, p. 153303382210853, 2022. https://doi.org/10.1177/15330338221085375

[21] J.Yang, L. Por,A. Khan, M. Khan, H. Karamti,A.Aldweeshet al., "Improving lungs cancer detection based on hybrid features and employing machinelearning techniques",, 2023. https://doi.org/10.21203/rs.3.rs-2929395/v1

[22] A. Akram and M. Iqbal, "Deeprnascanner: deep learning-based discovery of regulatorymirnasequencesinlungcancer",,2023.

https://doi.org/10.21203/rs.3.rs-3144904/v1

[23] S. Rai, D. S, J. Pan, D. Mishra, & X. Fu, "Multigroup prediction in lung cancer patientsandcomparativecontrolsusingsignatureofvolatileorganiccompounds inbreathsamples",PlosOne,vol.17,no.11,p.e0277431,2022.

https://doi.org/10.1371/journal.pone.0277431

[24] E. Klein, D. Richards, A. Cohn, M. Tummala, R. Lapham, D. Cosgroveet al., "Clinicalvalidationofatargetedmethylation-basedmulti-cancerearlydetection test using an independent validation set",Annals of Oncology, vol. 32, no. 9, p. 1167-1177, 2021. https://doi.org/10.1016/j.annonc.2021.05.806

[25] Y. Cui, X. Shi, S. Wang, Y. Qin, B. Wang, X. Cheet al., "Machine learning approaches for prediction of early death among lung cancer patients with bone metastases using routine clinical characteristics: an analysis of 19,887 patients", Frontiers in Public Health, vol. 10, 2022.

https://doi.org/10.3389/fpubh.2022.1019168

[26] T. Sajja, R. Devarapalli, & H. Kalluri, "Lung cancer detection based on ct scan images by using deep transferlearning",Traitement Du Signal,

vol. 36,no.4, p. 339-344, 2019. https://doi.org/10.18280/ts.360406

[27] R. Tekade, "Lung nodule detection and classification by using convolutional neural network", Helix, vol. 8, no. 5, p. 3696-3700, 2018. https://doi.org/10.29042/2018-3696-3700

[28] C. Huang, C. Zeng, Y. Wang, H. Peng, C. Lin, C. Changet al., "A study of diagnostic accuracy using a chemical sensor array and a machine learning technique to detect lung cancer", Sensors, vol. 18, no. 9, p. 2845, 2018. https://doi.org/10.3390/s18092845

[29] W. Choi, J. Oh, S. Riyahi, C. Liu, F. Jiang, W. Chenet al., "Radiomics analysis of pulmonary nodules in low-dose ct for early detection of lung cancer",Medical Physics, vol. 45, no. 4, p. 1537-1549, 2018. https://doi.org/10.1002/mp.12820

[30] Jemal A, Siegel R, Xu J, et al. Cancer statistics, 2010. CA Cancer J Clin. 2010;60(2):9-25. doi:10.3322/caacj.20078