iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** IV **Month of publication:** April 2025

DOI: https://doi.org/10.22214/ijraset.2025.68232

A Novel Machine Learning Framework for Predicting Lung Cancer from Clinical and Imaging Data

Lokesh S Wandhare[1], Pranay R Chippawar[2], Pranay V Tijare[3], Yash S Kosare[4], Girish B Umaratkar[5]
[1, 2, 3, 4]*Student,* [5]*Assistant Professor, Artificial Intelligence & Data Science Department of Priyadarshini College of Engineering, Nagpur, Maharashtra, India*

Abstract: Lung cancer is one of the most common and lethal cancers globally, and early detection plays a crucial role in improving patient survival. Traditional diagnostic techniques rely on manual analysis of medical images, which can be time-consuming and susceptible to human error. This paper discusses the use of AI tools in the early detection of lung cancer. This can be used to improve the detection accuracy significantly as against conventional methods, thereby allowing AI assistance to radiologists to make more accurate and timely diagnoses. The framework further allows it to be scaled and adapted for different imaging modalities to be implemented in real clinical settings. This research will show the transformative impacts that AI has on healthcare, especially against diseases like lung cancer, where early detection is key.
Keywords: Lung Cancer Detection, Machine Learning, Image Segmentation, CT Scan Image.

I. INTRODUCTION

Lung cancer, being among the top cancer killers around the world, has contributed to the ever-increasing need for early detection and accurate classification of lung abnormalities using advanced diagnostic tools. In the recent past, Transfer Learning, an aspect of AI and deep learning, has shown impressive results in detecting and diagnosing lung cancer from medical images. This result from such a study can now form the basis for an automation system that is not just capable of analysing images or lung cancer but will offer the promise of varied improvements in diagnostics, error control, and speedy evaluation to the medical world. The workflow of this system starts with the collection and curation of a lung cancer image dataset, which stands as the core of any machine learning model aimed at medical image analysis. The dataset would usually have several images of lungs, some normal ones and others containing cancerous abnormalities like nodules or tumour. It then applies preprocessing techniques like resizing, normalization, and augmentation to ensure uniformity in the scale, quality, and format of images, which in turn will improve the performance of the model. Augmentation techniques include rotation, flipping, and scaling, all of which are very effective in artificially increasing the diversity of the dataset, hence reducing the risk of overfitting and enabling better generalization when the model makes predictions on unseen data.

II. ISSUES AND PROBLEM

Lung cancer is one of the foremost causes of death due to the complaint, and it's nearly entirely because of late-stage discovery. traditional styles that are used to diagnose cancer through medical imaging are a time-consuming process and depend on mortal error.

Indeed, in early-stage lung cancers, nodes aren't fluently detected manually as the abnormalities are minor and bitty.

The lack of large, labelled medical image datasets limits the development of accurate AI models for lung cancer discovery.

Training CNNs from scratch on medical imaging tasks requires enormous quantities of data and computational coffers, which is hamstrung for lower datasets.

The accurate isolation between benign and nasty lung abnormalities is essential for proper treatment planning.

There's an adding demand for secure AI-driven systems able of aiding radiologists in achieving real-time prognostications of lung cancer from medical images through automated means.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue IV Apr 2025- Available at www.ijraset.com

III. METHODOLOGY

The process starts with inputting lung CT scan images, which are then preprocessed to improve clarity and reduce noise. This is achieved using a geometric mean filter, which smoothens the image without compromising important details, making it easier to identify relevant regions in later steps.

Once the image is processed, segmentation is carried out using the K-means clustering algorithm. This step divides the image into distinct sections, isolating potential tumor areas from the surrounding healthy lung tissues. Proper segmentation is critical, as it helps focus the analysis on the areas of interest where abnormalities may be present. After segmentation, important features are extracted from the image using Linear Discriminant Analysis (LDA), which reduces the complexity of the data by selecting the most significant characteristics that distinguish between benign and malignant tumors.

The final stage involves classification, where the extracted features are analyzed to predict whether the detected tumor is cancerous (malignant) or non-cancerous (benign). This entire workflow, from image preprocessing to tumor classification, is designed to streamline the detection process, making it faster, more accurate, and potentially life-saving by aiding early diagnosis and treatment planning for lung cancer patients.

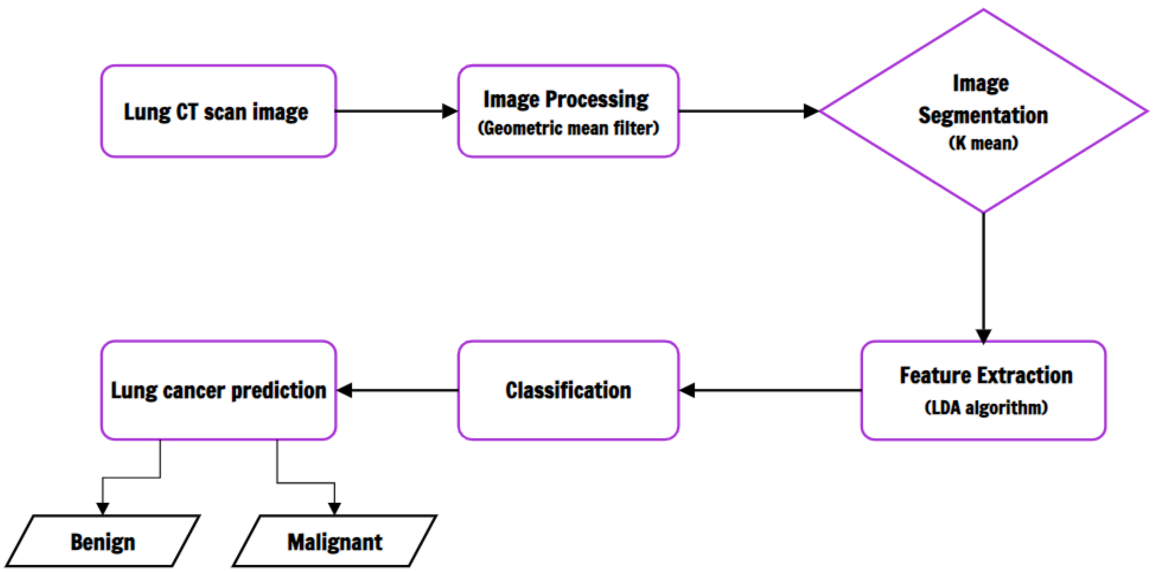


Figure 1: Proposed methodology for lung cancer detection using CT scan image

IV. PROPOSED LUNG ANALYSIS METHOD

It's concentrated on lung segmentation which is needed for the mainframe backed analysis from CT checkup images and it is to unnaturally divide the voxels original to the lung concave in the axial CT checkup separates from the conterminous lung deconstruction. We've proposed a scheme that first performs an most favourable thresholding which selects the threshold grounded on the reality and background pixel means. Once the threshold is chosen and functional, also the fiefdom mounting and connectivity disquisition is used to get the factual crater region with delicacy. The bracket begins by preprocessing and accelerating the image regions attained in the segmentation process. An ANN is also trained with back- propagation using the stoked dataset. It's achieved to reduce lapping findings through a conterminous image bracket and rejection rule.

A. Lung Nodule Verification with Ann and Hereditary Algorithm:

Neuro- Genetic Segmentation In sketching the division prove using Artificial Neural Networks, the model simplicity, generalizability and acceptability are essential to the robustness of the study to be carried out. Model simplicity depends upon the optimum way of hybridizing two different procedures in which only the simplest thesis that fits the fact of an issue is to be considered. In the current test Artificial Neural Networks and inheritable calculation with minimum parameters were used in the calculation. Meanwhile, representativeness of the show depended on the quality of the data which must represent the type of data that's being delved.

Both Artificial Neural Networks and inheritable Algorithm are protean, important and ready to bargain effectively with an expansive variety of issues including profoundly nonlinear models and uproarious information. also, they do not bear before data to display the issue being studied. thus, from a realistic point of view, Neural Network and inheritable Algorithm appeared to 53 workshops best in mix. The Neurogenetic approach performs programmed and hearty lung division by figuring the ideal edge of the picture.

B. The proposed fashion deals with the posterior way.

Preprocessing → Thresholding → Conditions reduction → Border discovery → Image segmentation

Lung segmentation is the same as mentioned over. Coming model is to be demonstrated cast the represented form of Background image filtering inheritable segmentation.

Lung area in CT cross sectional image: Lung dataset of frontal chest x-beams requested by the Japanese Society of Radiological Technology (JSRT) and Dataset of CT images were taken from the lung Image Database Consortium (LIDC) which have been utilized as the test images for the tests. The set comprised of 200 chest-beams, out of which 125 x-beams were anomalous and 81 x-beams were ordinary. It is aiming to state that the lungs in the X-beam had been affected by a few ailments that had been talked almost in segment 1. The LIDC dataset comprised of 200 CT pictures out of which 196 pictures were atypical and 102 were ordinary pictures. All x-beam pictures had a measure of 1024x1024 pixel sand dull scale shading significance of 12 bit. The pictures were isolated into two areas. One half was utilized in arrangement and the other half utilized in testing the calculation. Deciphering a chest radiograph was surprising work. Radiographs regularly involved impressive differentiation assortments, and basic moo complexity interface as appeared in the code below.

Image preprocessing in contaminated lung zone: Keeping up the extreme objective to portion the lung imaged successfully the analyst included the preprocessing step for the proposed strategy to decide whether the data picture was polluted with a few kinds of clamour. In case of the polluted picture, the steps of division may be affected. Picture improvement procedures improve pictures see. Picture improvement was one arrangement to a PC imaging issue. Various picture upgrade methods were emphasized to upgrade picture highlights for display and investigation. The histogram was gotten for the picture partitioning the centre between least and maximum pixel esteem into similarly isolated buckets. The analyst allotted each pixel to the bucket that contained its esteem. At that point, number of pixels comparing to each bucket was tallied. The histogram of this image repeats is considered portion of the repository range. This repeat considers states of the histograms for a comparative picture alter depending upon the measure of the interims. The histograms were the premise for numerical spatial region arrangement procedures (Brant et al., 2012). Histogram control was a productive method for picture enhancement. Histogram levelling out was one of the most vital parts of any picture preparing. The essential principle of histogram levelling was that all the picture powers ought to be similarly visit. A picture whose pixels had a propensity to have the entirety scope of conceivable dull levels and, moreover, had an inclination to be appropriated reliably would have an appearance of tall separation and show a tremendous collection of dim tones. Peaks in the histogram talked to visit pixel powers, and may as often as possible be distinguished with nearly homogeneous regions.

After histogram levelling out, the crests extended, meaning that humble control contrasts in a locale got to be steadier. Histogram adjusting did not "straighten" a histogram. It reallocated control streams.

Lung development classification: A conspire planned to convey a multi-scale neural organize engineering that would recognize normal handles from layout plans. A setup of Gabor channels, as well as a Laplacian of Gaussian channel is utilized for include end from the locales of data and to help a 3-layered neural network by utilizing MATLAB. Major of the plans separate ranges into handles to snatch highlights from a specific part of handles. At this organize, they convolve the data locales with a course of action of Gaussian channels to isolate estimations from the internal and band regions of the handle. A two-step lung picture division window order is carried out utilizing unforgiving k-closest neighbour calculation.

Image Thresholding and dismemberment: Image presently had superior separation in any case there was too much immaterial establishment information and mess that should have been evacuated. It was found that most of the foundation data by pixel values were not the same as those of the lungs. Picture thresholding was a subclass of picture division as its allotments a picture into parcels in see of the estimation of pixels with regard to restrain regard. Ideal thresholding was the starting stage in thresholding the picture. A thoracic CT contained two essential gatherings of pixels: tall control pixels arranged in the body and moo constrain pixels that were in the lung and the encompassing discuss. The tremendous contrast in drive between these two bunches driven to a great isolation since of thresholding. Thresholding was utilized for certification of the real twofold covers for the lung locale. It produces twofold shroud from the input dim level CT utilizing an iterative edge calculation, which is a way better method than the unique limit calculation where the limit was chosen basically as the least of the two maxima from the histogram of the dim level.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue IV Apr 2025- Available at www.ijraset.com

The picture histogram was to begin with fragmented into two parts utilizing a start restrain esteem, which may for case be half of the greatest of the dynamic extend of the current picture, or the traditional edge esteem fair depicted. After that, the test cruel of the dull highlights related with the closer view pixels and the test cruel of the dim highlights related with the base pixels were computed, and edge value was decided.

Image segmentation is another vital operation in image examination. It breaks an image into bunches of pixels that shape a locale of the image. Applications of 2D image division in the wellbeing division incorporate volume estimation and visual representation of objects of intrigued, location of anomalies, and tissue measurement and separation among others (Lee cum et al, 2015). The point of division is to change the image representation into something significant and more agreeable to elucidation. Image division is frequently utilized to discover objects and boundaries in images. In other words.

Image segmentation is the handle of partner a name to each pixel in a picture such that pixels with the same name have a few visual properties. It produces a set of comparative sections that collectively frame the total picture. All pixels in a given locale are comparative with respect to a few characteristic or computational property, such as colour, concentrated or texture.

Results on lung image background removal and segmentation: Base Removal After picture thresholding, base evacuation of the picture was done. By applying the constrain to the picture, the entirety lung picture might be obtained from the base. So, there was a require of a base departure instrument to remove the base. Hence the histogram-based technique was utilized for this reason

V. RESULT AND DISCUSSION

The results of the proposed system demonstrate its effectiveness in enhancing diagnostic images and detecting lung nodules with high precision. By utilizing the Multiresolution Rigid Registration (MRR) method, the system accurately aligns input CT images at varying resolution levels, ensuring that critical lung structures and nodules are preserved and correctly positioned. This alignment reduces misalignment errors, resulting in a composite image that is visually coherent and diagnostically reliable. The system's ability to maintain image integrity enhances the accuracy of medical evaluations.

The DWT-PCAv Fusion technique further improves the quality of these registered images by merging information from multiple image resolutions and decomposition levels. This fusion process enhances the visibility of subtle diagnostic features, such as small nodules and vascular structures, which are crucial for detecting complex lung conditions. The resulting images offer richer, more detailed information, aiding radiologists in accurately identifying abnormalities and improving the precision of early-stage diagnoses.

In addition to image enhancement, the system excels in detecting and classifying lung nodules. The enhanced images enable the identification of potential abnormalities in the lung parenchyma with high sensitivity, minimizing the risk of overlooking even the smallest nodules. Moreover, the classification system effectively distinguishes between benign and malignant nodules with impressive accuracy, reducing false positives while maintaining a low false-negative rate. These advancements highlight the system’s potential as a valuable tool for improving diagnostic workflows and patient outcomes in clinical radiology.

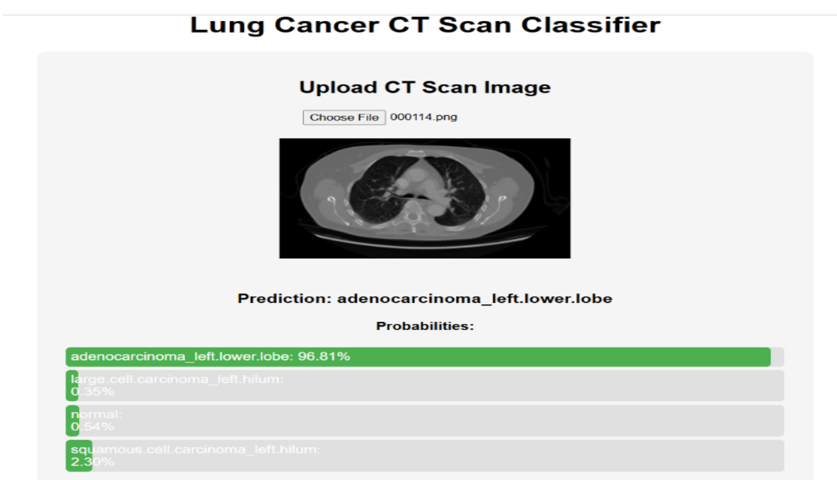


Figure 2: "adenocarcinoma left lower lobe" with a high confidence level (96.81%) along with other probabilities for different classes

VI. CONCLUSION

This paper reviews various steps of lung cancer detection and formation of the lung cancer detection system using Transfer Learning, illustrating the high potential of AI in improving early diagnosis of lung cancer. With high accuracy, sensitivity, and specificity, the system represents a promising tool for supporting healthcare professionals in making informed decisions. While limitations still exist, furthering work to overcome these will lead to an evolution of AI-based diagnostic tools in clinical medicine towards improved management and patient care for lung cancer.

VII. FUTURE SCOPE

The future of lung cancer detection using machine learning looks incredibly promising, with advancements that could make diagnoses even faster, more accurate, and widely accessible. As AI models continue to improve, they will become even better at spotting early signs of lung cancer, potentially detecting tiny abnormalities that might be overlooked by the human eye. With the growing use of deep learning and improved image processing techniques, future systems could not only identify cancer but also predict its progression, helping doctors personalize treatment plans for each patient.

Another exciting possibility is the combination of AI with other medical technologies, such as genomics and biomarker analysis, to create a more comprehensive approach to lung cancer diagnosis. by analysing genetic and molecular data alongside imaging scans, AI could help doctors determine the best treatment options based on a patient's unique profile. Furthermore, as AI becomes more trusted in the medical field, its use could expand to underserved areas where access to specialized radiologists is limited, bringing high-quality cancer screening to more people worldwide.

VIII. ACKNOWLEDGEMENT

We would like to take this opportunity to thank our faculty mentors and project guides for their constant support, encouragement and directions which greatly contributed to the completion of this study. We are grateful to the Department of Artificial Intelligence & Data ScienceDepartment of PriyadarshiniCollege of Engineering, Nagpur, for their support of resources and facilities for this project. We also extend our gratitude to our families and friends who have supported us throughout the process of this work. This work has been a team effort and we are grateful to any individual who has been there to support us in any way throughout this process.

REFERENCES

[1] Nall R. What to Know about Lung Cancer. Medical News Today. 2018. [(accessed on 2 April 2022)]. Available online: https://www.medicalnewstoday.com/articles/323701

[2] Saeed S., Jhanjhi N.Z., Naqvi M., Humyun M., Ahmad M., Gaur L. Approaches and Applications of Deep Learning in Virtual Medical Care. IGI Global; Harrisburg, PA, USA: 2022. Optimized Breast Cancer Premature Detection Method with Computational Segmentation: A Systematic Review Mapping; pp. 24–51.

[3] Facts About Lung Cancer. [(accessed on 27 March 2022)]. Available online: https://www.lungcancerresearchfoundation.org/lung-cancer-facts/

[4] Shukla D. New Test May Quickly Detect Early-Stage Lung Cancer. Medical News Today. 2022. [(accessed on 28 March 2022)]. Available online: https://www.medicalnewstoday.com/articles/new-test-may-quickly-detect-earlystage-lung-cancer

[5] Zhou Y., Lu Y., Pei Z. Accurate diagnosis of early lung cancer based on the Machine Learning model of the embedded medical system. Microprocess. Microsyst. 2021;81:103754. doi: 10.1016/j.micpro.2020.103754.

[6] Nanglia P., Kumar S., Mahajan A.N., Singh P., Rathee D. A hybrid algorithm for lung cancer classification using SVM and Neural Networks. ICT Express. 2021;7:335–341. doi: 10.1016/j.icte.2020.06.007.

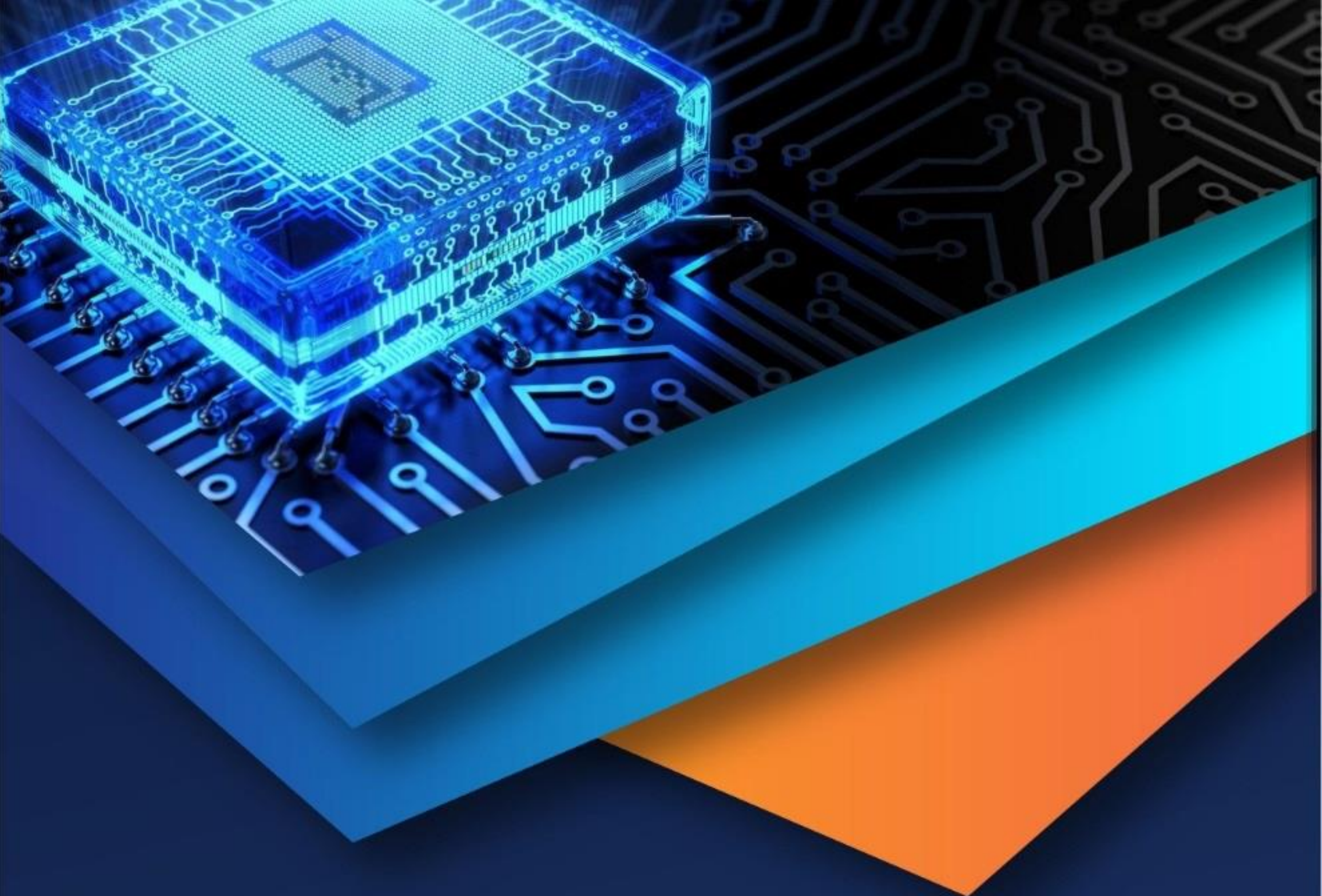
[7] Bai Y., Li D., Duan Q., Chen X. Analysis of high-resolution reconstruction of medical images based on deep convolutional neural networks in lung cancer diagnostics. Comput. Methods Programs Biomed. 2022;217:106592. doi: 10.1016/j.cmpb.2021.106592.

[8] Ahmad A.S., Mayya A.M. A new tool to predict lung cancer based on risk factors. Heliyon. 2020;6:e03402. doi: 10.1016/j.heliyon.2020.e03402.

[9] Ahn B.-C., So J.-W., Synn C.-B., Kim T.H., Kim J.H., Byeon Y., Kim Y.S., Heo S.G., Yang S.D., Yun M.R., et al. Clinical decision support algorithm based on machine learning to assess the clinical response to anti–programmed death-1 therapy in patients with non–small-cell lung cancer. Eur. J. Cancer. 2021;153:179–189. doi: 10.1016/j.ejca.2021.05.019.

[10] Agarwal S., Yadav A.S., Dinesh V., Vatsav K.S.S., Prakash K.S.S., Jaiswal S. By artificial intelligence algorithms and machine learning models to diagnose cancer. Mater. Today Proc. 2021 doi: 10.1016/j.matpr.2021.07.088.

[11] Huang X., Lei Q., Xie T., Zhang Y., Hu Z., Zhou Q. Deep transfer convolutional neural network and extreme learning machine for lung nodule diagnosis on CT images. Knowl.-Based Syst. 2020;204:106230. doi: 10.1016/j.knosys.2020.106230.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY