Volume: 13 Issue: I Month of publication: January 2025

DOI: https://doi.org/10.22214/ijraset.2025.66770

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 I Jan 2025- Available at www.ijraset.com

Prediction of Lung Cancer Using Machine Learning

Avinash S[1], Anu Kiruthiga R[2], Aditya Saran K P[3], Harish A[4], Dr. D. Rasi[5]
Department of Computer Science and Engineering, Sri Krishna College of Engineering and Technology, Coimbatore, India

Abstract: With the highest cancer-related mortality, lung cancer is one of the most dangerous and difficult cancers to identify. Over time, machine learning techniques have been used in medical research, and many significant developments have been made in cancer research for decades. There have been a lot of machine learning algorithms that could significantly help in feature extraction and other radiological analysis. With the help of available datasets from research organizations, it is made possible easily. In this study, we built a system with the help of the YOLO algorithm that uses CNN to predict lung cancer from the fed datasets of cancerous and non-cancerous CT scans.

Keywords: Mortality, Machine learning algorithms, Feature extraction, Datasets.

I. INTRODUCTION

With nearly 10 million deaths by the year 2020 reported by the WHO, cancer tends to be a menacing disease. In the year 2020, it was found that 1.79 million people were dead due to lung cancer. Which tops the mortality rate of any other cancer types. It has been found that This is the most difficult cancer to identify since symptoms appear only in the later stages and are negligible or non-existent in the early stages. Yet, early detection and diagnosis have been shown to be vital for survival.

Cancer cells are generally defined as cells that proliferate uncontrollably, leading to the creation of tumour. This process, known as metastasis, allows the cancerous cells to spread from one area of the body to another through blood and lymphatic vessels. Genes that have mutated in tumour cause normal cells to become cancerous ones. Over the period, many different approaches have been discovered in predicting Lung Cancer using machine learning, many of those seems to give desirable outcomes. Some of the most commonly used machine learning and deep learning algorithms are support vector machines (SVM), decision trees, random forests, K-neighbors, artificial neural networks (ANN), and convolutional neural networks (CNN). Smoking is one of the leading causes of lung cancer; even people who do not smoke have been impacted by lung cancer. Such other factors found were exposure to harmful gases, bad living conditions, air pollution and in some rare cases it can also be inherited as it is found to a genetic disease.

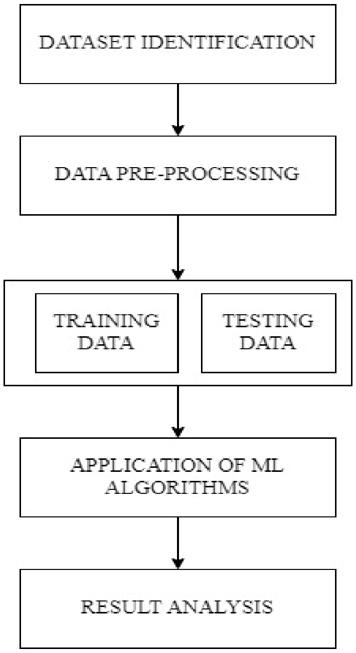


Fig.1 General working flow of cancer prediction systems

II. RELATED WORKS

Lung cancer diagnosis has a lengthy record. The examination of tumour markers and pathological findings have grown in importance as diagnostic techniques. Both methods, however, have certain drawbacks. For example, invasive treatments are the gold standard for diagnosing abnormal results. However, other indicators, such as carcinoembryonic antigen, have low specificity, which could result in inaccurate clinical diagnoses. Diagnostic imaging methods like CT, MRI, and CXR can help with the diagnosis of lung cancer. Yet, certain tiny lung nodules or lymph node metastases can still be missed or difficult to detect radiographically. Thus, AI became a useful technique for diagnosing lung cancer.

Even with those outcomes generated from the inputs fed to the system by these algorithms and all other different approaches, it still requires pathologist help in the work flow as well as communication of results to the patients and directing them to an appropriate physician. Which can be overcame with the help of Natural Language Processing (NLP). With the advancements we have with us at present it might not be achieved, but in near future it can be achieved by creating a system which gets a patient's pathological findings and impact analysis as inputs which generates a textual result consisting the diagnosis of the findings, which can then be used to make a proper analysis of the prior pathologist's work.

In general, a system which uses machine learning to predict cancer, works with a common flow (Fig. 1): The dataset for the system is identified, they can be of clinical data as well as radiological reports which depends on the methodology used in the system. After finding a dataset which would work for the prediction, they are pre-processed to make a best fit for the system, after being processed the data are then split in two individual groups one for training the system and the other for testing the system, with this almost every part of the work is done. Once after that Machine Learning algorithms are used for predicting outcomes which are then used to evaluate the efficacy and accuracy. ***[3]***

The study aims to explore radiomic classifiers for lung cancer histology. In this study, two independent radiomic cohorts were analyzed, and 440 radiomic features were extracted from permanent CT images. Univariate analysis identified 53 features significantly associated with tumor histology; multivariate analysis revealed Naïve Bayes as the best classifier. The study finds that radiomics combined with machine learning can predict lung cancer histology, and Naïve Bayes achieved high accuracy in classification. ***[4]***

In this study, they evaluated the performance of two machine learning algorithms, SVM and LR, in predicting the survival rates of lung cancer patients. The effectiveness of these algorithms is compared based on accuracy, precision, recall, F1 score, and confusion matrix. In the end, logistic regression seems to outperform SVM in terms of accuracy.***[5]***

The study aims to improve lung nodule detection using a hybrid feature set and ANN. In this study, lung volume is extracted from input CT images using optimal thresholding, and image enhancement is performed using multi-scale dot augmentation filtering. Lung nodule candidates are detected in the enhanced image. Features are extracted, including texture, 2D and 3D shapes, and intensity. A two-layer feed-forward neural network classifies lung nodules based on extracted features.***[6]***

The study's goal is to improve lung cancer detection and prediction by employing an intelligent computer-aided diagnosis system. The suggested approach uses multi-stage classification to identify lung cancer. Segmentation and picture enhancement are carried out independently at each stage. To achieve precise tumor localization, methods including thresholding and marker-controlled watershed segmentation are employed. Other techniques include picture scaling, color space transformation, and contrast enhancements. The system predicts and detects lung cancer using a multi-class SVM classifier. The approach shows improved lung cancer detection and prediction accuracy.***[7]***

The study employs machine learning algorithms to predict early stages of lung cancer. In addition to a deep learning algorithm called Artificial Neural Network, four machine learning algorithms—Bayes Net, Naïve Bayes, Decision Tree, and Random Forest—are taken into consideration for the study. The lung cancer dataset measures and assesses cutting-edge factors. According to the study, ANN performs better than other algorithms and is the best learning algorithm overall due to its increased accuracy.***[8]***

This study compares several classification algorithms, including Naïve Bayes, SVM, Decision Trees, and Logistic Regression, using training data from Data World and UCI. The accuracy of each model is determined by evaluating the results, and the algorithms' performances are looked at to determine which is the most efficient. Using the k-fold cross validation procedure, the dataset is divided.***[9]***

The study discusses how important it is to find lung cancer early on. The suggested approach combines a self-normalized multi-view convolution neural network with adaptive boosting. This approach uses several perspectives of the lung nodules to achieve reliable detection. It is discovered that while Multi View CNN (MV-CNN) can function with numerous inputs, CNN can only process one input at a time. Within the suggested methodology, MV-CNN functions as the baseline learner. Network layers are normalized by the scaled exponential linear unit activation function, and regularization is improved using a unique dropout strategy.

The Early Lung Cancer Action Program (ELCAP) and the Lung Image Database Consortium and Database Resource Initiative (LIDC-IDRI) datasets are used to train and evaluate the suggested approach. It is discovered that for lung nodule detection, SNMV-CNN achieves good accuracy, sensitivity, and specificity. AdaBoost-SNMV-CNN has the potential to be a non-invasive clinical diagnostic tool for lung cancer; its accuracy helps radiologists diagnose the disease early and furthers the cause of early lung cancer detection.***[10]***

In the proposed system few different smaller datasets are taken, the X-ray images are then pre-processed which include,

- Increasing contrast of the image using histogram equalization.
- Removal of noise using median filtering.
- Image resizing
- Image normalization based on mean standard deviation.

DenseNet-121 of CNN architecture is used for the proposed system, with this a base model is trained and subdivided into two modes further which are trained with two datasets and are evaluated individually, further the model is re-trained with another dataset too.

III. COMPARISON CHART

Paper by	Year published	Algorithm used	Accuracy	Split ratio
Akram et al. [5]	2015	ANN	96.68%	Tr: 50% Ts: 25%
Wu et al. [3]	2016	Naïve Bayes	72%	Tr: 198 samples, Ts: 152 samples
Hazra et al. [4]	2017	Logistic Regression	77.4%	Tr: 80% Ts:20%
Worawate et al. [10]	2018	DCNN	84.02%	Tr: 80% Ts: 10%
Janee Alam et al. [6]	2018	SVM	87%	Random
Radhika et al. [8]	2019	SVM	99.2%	Split by k-fold cross validation technique
Deepak Rawat et al. [7]	2022	ANN	92.23%	Split k-fold cross validation technique
Adeel khan et al. [9]	2022	AdaBoost-SNMV-CNN	92%	Tr: 80% Ts: 20%

In the above table following acronyms have been used:
Tr- Training data
Ts- Testing data
ANN- Artificial Neural Network
DCNN- DenseNet Convolution Neural Network
SVM- Support Vector Machine
AdaBoost- Adaptive Boost
SNMV- Self Normalized Multi-View
CNN- Convolution Neural Network

IV. PROPOSED SYSTEM

In our system we utilize CNN, based on lung CT scan classifications. We pre-processed the lung cancer detection dataset consisting of cancerous and non-cancerous CT scans. With help YOLO v8 classification model we trained a model using our own customized dataset. In the beginning dataset is classified into 3 classes: benign, malignant and normal. From the classes made dataset is further divided into 3 subfolders for training, testing and validation. With these data a model is developed for the custom dataset.

1) *Data acquisition:* Datasets of lung CT scans used for cancer detection are identified and gathered from oncologist verified data.
2) *Data preparation:* After acquiring the dataset, the data are classified as normal, benign and malignant. Which are further split into train, test and validate.

3) *Yolov8n-class model:* A model created for classification of different objects from the COCO dataset, with help of this model a custom model is derived.
4) *Best:* While training the yolov8n-class model with our own dataset the parameters identified for best classification results are found and appended to another model called best, this model holds the highest matching parameters which could identify the classifications from our dataset.
5) *Validation data:* From the data split as train, test and validate, the data on the validation is sent over the model derived to check with the model's accuracy on customised dataset.
6) *Nodule detection:* With the validation data sent to the derived model the system identifies nodules on the CT images.
7) *Classified results:* After detecting the nodules the system compares the nodule size with that of the trained and tested data for classifying them whether the detected nodule is a benign or a malignant.

The dataset is split as 70% for training, 10% for testing and 20% for validation.

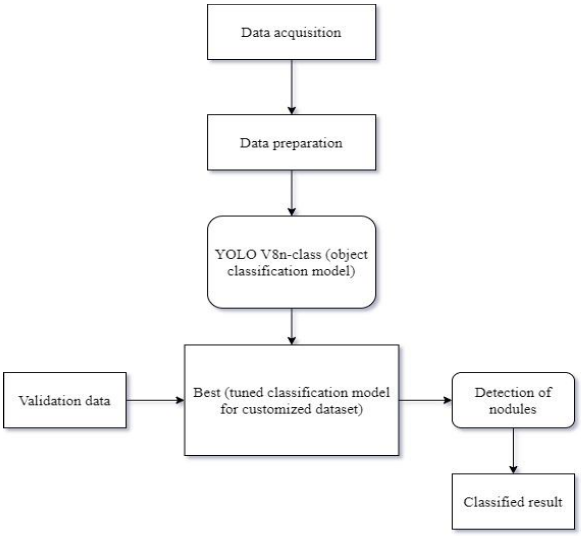

Data acquisition
Data preparation
YOLO V8n-class (object classification model)
Validation data
Best (tuned classification model for customized dataset)
Detection of nodules
Classified result

Fig.2 Working of the system.

V. RESULT AND ANALYSIS

It is found that the overall accuracy of the model on training and evaluating the data provided is comparatively higher than the similar hybrid CNN model used in the system proposed by Khan A et.al [9].

Fig.3 suggests that: train/loss, a measure of how well the model is fitting the training data shows that the model is learning from the training data. val/loss, a measure of how well the model is performing on unseen data shows the model performs well with data used like the training data, and can be increased with increases in amount of data. metrics/accuracy top1, is the percentage of times that the model predicts the correct class for a given input shows that the model is learning to classify the data more accurately. metrics/accuracy top5, is the percentage of times that the correct class is one of the top five predictions of the model. This shows the top-5 accuracy is also increasing over time, but it is higher than the top-1 accuracy. This is because it is easier for the model to predict one of the correct five classes than it is to predict the exact correct class.

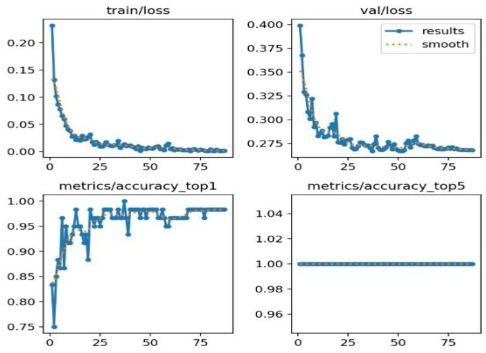


Fig.3 Performance graphs

Overall, these graphs suggest that the model is learning and performing well. The training loss and validation loss are both decreasing, and the accuracy is increasing. However, it is important to note that the validation loss is still higher than the training loss, which suggests that the model may be overfitting to the training data. This means that the model is learning patterns that are specific to the training data.

Fig.4 is a confusion matrix that describes the performance of a model in classifying between classes: malignant, benign and normal. Overall, the model seems to perform well, with an accuracy of 95.24%.

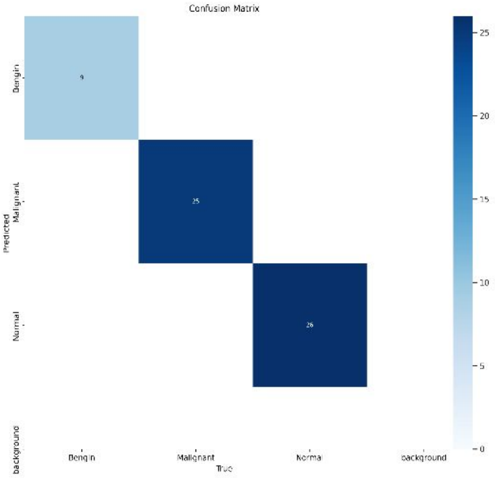


Fig.4 Confusion matrix

Class	Precision	Recall	F1-Score
Benign	1	1	1
Malignant	0.93	0.96	0.94
Normal	0.96	0.93	0.95

Table.1 Scores

Inferences made with references to the Table.1 and Fig.4, the proposed model is particularly good at identifying malignant samples, with a recall of 0.96. Which is important as misclassifying a malignant sample as benign could lead to a delay in diagnosis and treatment. The model is also good at identifying benign samples, with a precision of 1.00. Which is important as misclassifying a benign sample as malignant could lead to unnecessary anxiety and medical procedures. The model is slightly less accurate at identifying normal samples, with a recall of 0.93 and a precision of 0.96.

VI. FUTURE SCOPE

Real-time object detection is the goal of the majority of robotics and computer vision systems. Because of the early research in this area, it is creating output in a range of directions with significant advancement. In many scenarios, the Yolo algorithm's object detection is underutilized, a problem that may be resolved in the future. In fact, in the domains of pattern recognition and computer vision, YOLO object detection in photographs has received a lot of attention recently. These processes, however still in the experimental phase, have the potential to relieve humans of repetitive chores that are better left to computers and other systems. Subsequent studies ought to concentrate on improving precision, handling complex scenarios, tracking many objects, and identifying objects particular to a certain domain.

VII. CONCLUSION

It's critical to diagnose lung cancer as soon as possible. This study compares the effectiveness of the most popular deep learning and machine learning algorithms for lung cancer prediction. A system that was designed to find ways to improve accuracy was assessed using performance metrics in order to analyze the results. As a result, the system has greater accuracy than the current ones.

REFERENCES

[1] G. Sruthi, C. L. Ram, M. K. Sai, B. P. Singh, N. Majhotra and N. Sharma, "Cancer Prediction using Machine Learning," 2022 2nd International Conference on Innovative Practices in Technology and Management (ICIPTM), Gautam Buddha Nagar, India, 2022, pp. 217-221, doi: 10.1109/ICIPTM54933.2022.9754059.

[2] D. Manivannan., N. K. Manikandan and M. Kavitha, "Discovering Lung Cancer Cell Using Machine Learning," 2022 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI), Chennai, India, 2022, pp. 1-3, doi: 10.1109/ICDSAAI55433.2022.10028961.

[3] Wu W, Parmar C, Grossmann P, Quackenbush J, Lambin P, Bussink J, Mak R, Aerts HJ. "Exploratory Study to Identify Radiomics Classifiers for Lung Cancer Histology". Front Oncol. 2016 Mar 30;6:71. doi: 10.3389/fonc.2016.00071.

[4] Hazra, Animesh & Bera, Nanigopal & Mandal, Avijit. (2017). Predicting Lung Cancer Survivability using SVM and Logistic Regression Algorithms. International Journal of Computer Applications. 174. 19-24. 10.5120/ijca2017915325.

[5] Akram, Sheeraz & Javed, Muhammad & Qamar, Usman & Khanum, Aasia & Hassan, Ali. (2014). "ANN based Classification of Lungs Nodule using Hybrid Features from Computerized Tomographic Images". Applied Mathematics & Information Sciences. 9. 183-195. 10.12785/amis/010124.

[6] J. Alam, S. Alam and A. Hossan, "Multi-Stage Lung Cancer Detection and Prediction Using Multi-class SVM Classifie," 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), Rajshahi, Bangladesh, 2018, pp. 1-4, doi: 10.1109/IC4ME2.2018.8465593.

[7] D. Rawat, Meenakshi, L. Pawar, G. Bathla, and R. Kant, "Optimized Deep Learning Model for Lung Cancer Prediction Using Artificial Neural Networks Algorithm," 2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, pp. 889-894, doi: 10.1109/ICESC54411.2022.9885607.

[8] R. P.R., R. A. S. Nair and V. G., "A Comparative Study of Lung Cancer Detection using Machine Learning Algorithms," 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), Coimbatore, India, 2019, pp. 1-4, doi: 10.1109/ICECCT.2019.8869001.

[9] Khan A, Tariq I, Khan H, Khan SU, He N, Zhiyang L, Raza F. "Lung Cancer Nodules Detection via an Adaptive Boosting Algorithm Based on SNMV CNN". J Oncol. 2022 Sep 26;2022:5682451. doi: 10.1155/2022/5682451.

[10] Ausawalaithong, Worawate & Thirach, Arjaree & Marukatat, Sanparith & Wilaiprasitporn, Theerawit. (2018). Automatic Lung Cancer Prediction from Chest X-ray Images Using the Deep Learning Approach. 1-5. 10.1109/BMEiCON.2018.8609997.

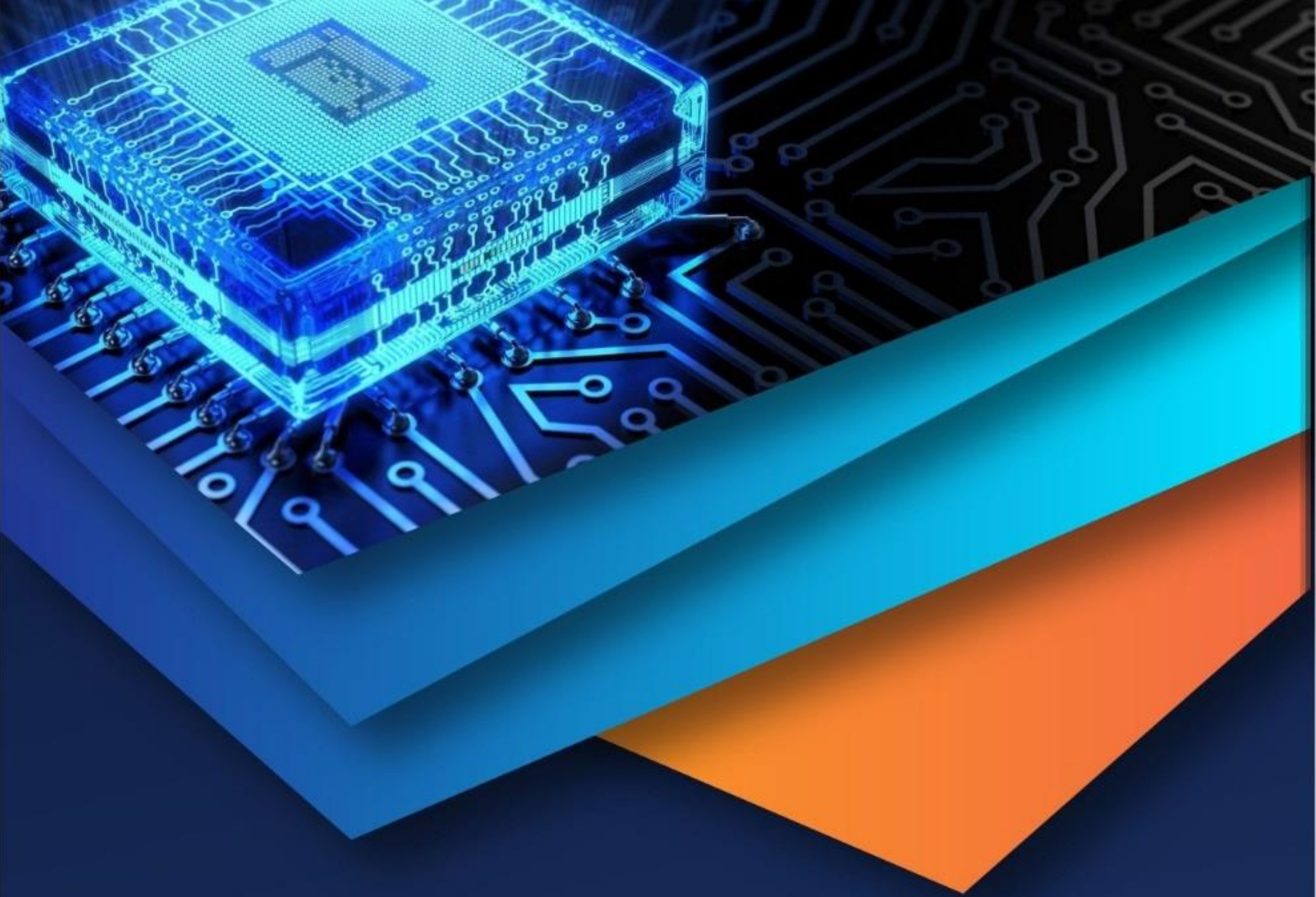
[11] V. Nisha Jenipher and S. Radhika, "A Study on Early Prediction of Lung Cancer Using Machine Learning Techniques," 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), Thoothukudi, India, 2020, pp. 911-916, doi: 10.1109/ICISS49785.2020.9316064.

[12] J. Al-Tawalbeh, B. Alshargawi, H. Alquran, W. Al-Azzawi, W. A. Mustafa, and A. Alkhayyat, "Classification of Lung Cancer by Using Machine Learning Algorithms" 2022 5th International Conference on Engineering Technology and its Applications (IICETA), Al-Najaf, Iraq, pp. 528-531, doi: 10.1109/IICETA54559.2022.9888332.

[13] J. Oentoro, R. Prahastya, R. Pratama, M. S. Kom and M. Fajar, "Machine Learning Implementation in Lung Cancer Prediction - A Systematic Literature Review," 2023 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Bali, Indonesia, 2023, pp. 435-439, doi: 10.1109/ICAIIC57133.2023.10067128.

[14] Huang S, Yang J, Shen N, Xu Q, Zhao Q. Artificial intelligence in lung cancer diagnosis and prognosis: Current application and future perspective. Semin Cancer Biol. 2023 Feb;89:30-37. doi: 10.1016/j.semcancer.2023.01.006.

[15] J. Chen, "Comparative Analysis of Machine Learning Models for Lung Cancer Prediction," 2023 IEEE International Conference on Image Processing and Computer Applications (ICIPCA), Changchun, China, 2023, pp. 242-246, doi: 10.1109/ICIPCA59209.2023.10257778.



10.22214/IJRASET

INDEX COPERNICUS

45.98

IMPACT FACTOR:
7.129

TOGETHER WE REACH THE GOAL

IMPACT FACTOR:
7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY