# Cancer Detection Using Machine Learning

B.Ranjitha[1], Shreetirth Talpallikar[2], SSS Anuroop[3]
[1]Assistant Professor, [2,3]UG Scholar
Department of CSE, Guru Nanak Institute of Technology, Hyderabad, Telangana, India
Email ID:ranjithab.csegnit@gniindia.org[1],talpallikar.shreetirth@gmail.com[2],
shanmukhanuroop444@gmail.com[3]

Citation
B. Ranjitha, Shreetirth Talpallikar, SSS Anuroop, "Cancer Detection Using Machine Learning", Journal of Next Generation Technology,5(6), pp. 43-49, October 2025.

**Abstract**
Cancer is still a mystery to many researchers and doctors working in medicine. Even today, the cause of cancer is unknown, and doctors have a hard time discovering it. This research aims to look into how technologies, such as "CNNs", "Deep Learning", and "Machine Learning", help in finding lung cancer. CNNs are part of Deep Learning models and are used for analyzing images. Using advanced technology, X-rays and CT pictures are checked to find spots or tumors in the concerned regions that lead to lung cancer. In this research, scientists use the IQ-OTH/NCCD-Lung Cancer Dataset" Molecular noise signals and scanned pictures are intuited by science to recognize cells and phases in the development of cancer. It relies mainly on data, algorithms, and the result of scientific studies in genetics and digital technology to find cancer in its early stages. It is really a form of solid knowledge, not only a labeling of symptoms. It allows doctors to prevent problems and provides hope and extra time to the patients. As a result, early detection becomes our most effective method of fighting against disease inside our bodies.

***Keywords: Convolutional neural networks, CT images, artificial intelligence, machine learning, lung cancer***

## I. INTRODUCTION

AI and machine learning have grown rapidly, which has caught the focus of researchers and doctors these days. Artificial intelligence in computer science aims to construct computers that are capable of tasks that need some human-like intelligence. Computers are able to learn from data on their own with machine learning (ML), which is a subfield of artificial intelligence, as long as they are not explicitly programmed. An interaction between the model and its environment, plus feedback, help it improve its understanding.

It makes use of techniques such as unsupervised learning, which searches for patterns not based on labels, and supervised learning, in which the models rely on supervision. The more refined area of machine learning known as deep learning utilizes artificial neural networks made up of many layers to deal with tough jobs such as recognizing speech and pictures.
Each year, lung cancer causes the death of millions, making it one of the most common and serious killers[1]-[4]. In 2020, there were over 1.8 million deaths due to lung cancer,

according to WHO. If lung cells begin to multiply out of control, they can build up tumours, disrupt breathing, and influence the body's normal working. Being a smoker greatly increases your risk of lung cancer, though the disease can happen to anyone. Lung cancer stays a dangerous issue, and most people with the disease do not recover, despite advancements in medicine and treatment.

Cancer is a mysterious and rather silent enemy that develops inside and disrupts the regular function of our cells. Other signs, such as genetic problems, unusual tissue growth, and buried chemical signals in blood and tissues, appear first before cancer is visible as symptoms. Detecting cancer means translating these faint and noisy signals in the body into coherent designs. Now, the detection of eye diseases can be made earlier using more techniques. It is changing into a mix of data-based models, genetics, using images in medical practices, and AI. Machine learning algorithms can now spot issues in x-rays with a much higher degree of accuracy. Liquid biopsies and analyzing biomarkers make it possible to detect cell changes well before tumours emerge[5]-[8].

The goal of education is clear yet very important: to change from waiting until things are serious to catching problems in advance and forecasting outcomes. Besides helping patients live longer, early diagnosis also enhances their quality of life, makes treatment less difficult, and provides them with personalised attention. It becomes a matter of thinking deeply about what is not immediately visible, trying to access things that cannot be read, and using compassion and imagination in their work[9]-[12].

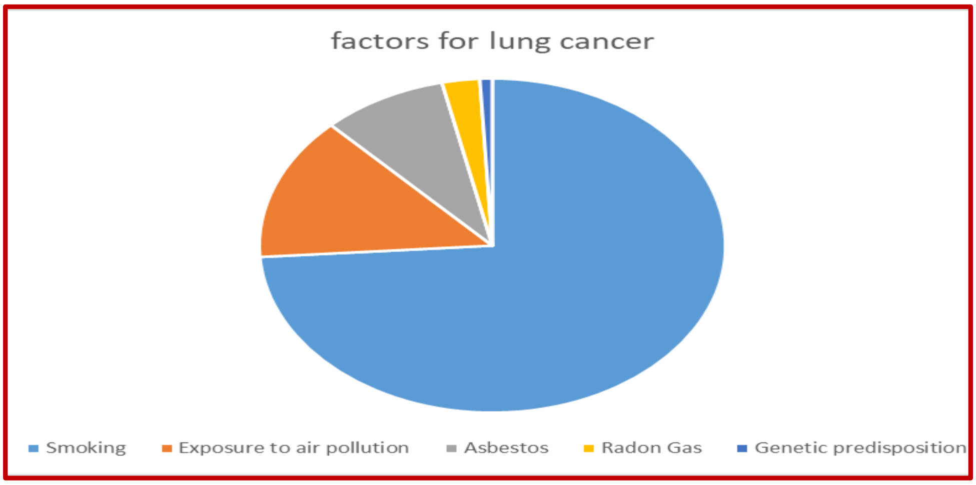

**Fig 1: Factors of Lung Cancer**

We applied pictures that were captured from several angles of "lung CT scans from the IQ-OTH/NCCD" tests for Cancer, which are carried out in a pathology facility. Usually, doctors depend on biopsies and CT scans for diagnostic purposes. Our model was trained using the Lung Cancer Dataset, which consists of high-quality lung images created by using a CT scan on patients.

## II. METHODOLOGY

### *A. Image Dataset Description:*

They use the IQ-OTH/NCCD lung cancer dataset, a set of CT scans that anyone can access and that have been chosen specifically for lung cancer studies. A total of 1,097 CT scan pictures from the year 2019 are included in the dataset, collected from 110 patients at the National Centre for Cancer Diseases/Iraq-Oncology Teaching Hospital. The goal of this dataset is to support the building of computer systems for diagnosing diseases by providing

real, annotated clinical data from professionals. Researchers are able to tell apart pulmonary nodule categories and healthy tissue using the carefully structured data, which has three separate categories of lung disorders. It is highly likely that all three groups contain benign lung nodules that are not cancerous. Malignant: Lung nodules that are cancerous.
One hundred and twenty fifties of the photos involve benign cases, another forty contains malignant cases, and the rest involve fifty-five normal cases. Being composed of many patient photo collections, the dataset is great for both segmenting and classifying medical images. Since both healthy lungs and all disease conditions have been included in the dataset, models will be able to tell the difference between them.

“The CT pictures were stored in DICOM format, which is the software standard used in the medical imaging industry. Photos were given the JPG format, and kept in their original size of 512 × 512 pixels to enable easy researching and analyzing. All the scans were taken with a Siemens SOMATOM CT scanner operating at 120 kV. The settings on the machine were fine-tuned to clearly image the lung tissues. A special setting of 350 to 1200 for window width and 50 to 600 for window centre was used to give the best contrast and visibility for soft tissues. Thanks to the 1 mm slice thickness always being maintained, high-quality imaging in thin sections was made easy for finding small nodules or mistakes.

The dataset was categorized and analysed by a team of experienced radiologists and oncologists, from the same medical centre. Experts have labelled these photos so well that they can be safely used for both training and testing AI systems in clinical work. Some of the ways people use this dataset include deep learning cancer diagnosis, extraction of features from images, and category identification for images[13]-[16]. All in all, the “IQ-OTH/NCCD dataset is a valuable tool for scientists focused on improving AI for early lung cancer detection. It is ideal for use in studies, comparisons, and tool-making due to its wide and diverse content, professional labels, and focus on medicine.

### *B. Areas of interest*

Each patient’s lung region of interest on the CT images was cut out and separated using regular morphological procedures and cropping systems. At the beginning, a binary lung mask that fitted these conditions was formed by using the thresholding method.

If “(x, y) > T, then G(x, y) = 1. 0, if (x, y) < T “

Here, x and y represent where a pixel is in the x and y directions, and G is the function that performs the thresholding operation. To indicate which area is the lung, it is given a value of 1, and the non-lung or background is marked with a value of 0. The value T divides lung tissue from its neighboring tissues. Where the pixel value is 1 in this binary mask, it means the pixel is part of the lung. Where it is 0, it shows the pixel is not part of the lung but is part of the ribs, chest wall, or organs around the lung. Removing unnecessary parts of the image forces the attention of the viewers to the main points.

The next step involved using dilatation, erosion, opening, and closing to clean up the mask of the lung. They make sure the segmented part is smooth, patch up small holes in the lung area, and remove minor particles of noise. Removing isolated false alarms or mistakes in the borders of the lungs is important, but the borders ought to remain the same. Fig. 2 shows the computed tomography (CT) scans of the lungs.

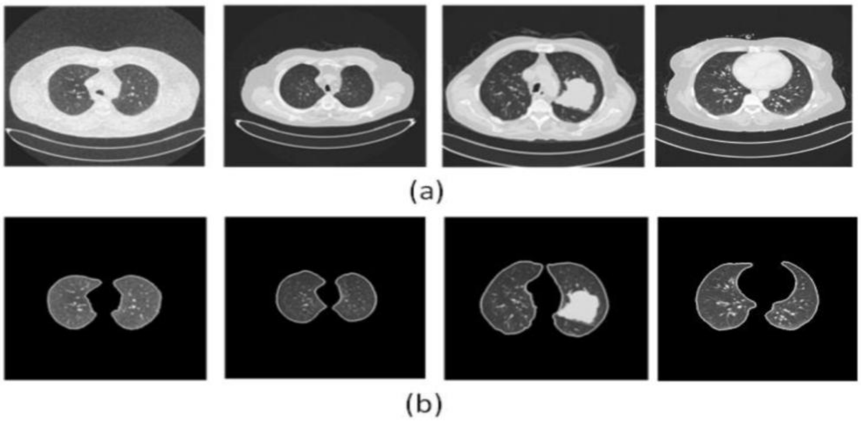

Fig. 2. computed tomography (CT) scans of the lungs

Before applying the cropping methods, the binary lung mask was fine-tuned to get the coordinates of the lung regions' bounding box. They extract the AOI from the CT scan by finding the smallest rectangle that holds all of the lung region. As a result, less unnecessary information is added, so more emphasis is placed on the meaningful part of the scan when it is studied and organised. Focusing training only on important parts of the lungs improves the results of downstream machine learning applications and also means the network has to analyze only relevant images. Being extra attentive during these early stages and while deciding if a nodule is benign or malignant is most important.

### *C. Data augmentation*

Data augmentation was used to add more data and strengthen diversity in the set. We used three ways to augment our data: Clients had the option to rotate their pictures at different angles, or use the vertical flip to turn them upside down, or the horizontal flip to turn them side to side. In addition, to fix the class imbalance and avoid it harming the results, the number of benign samples was made larger. Labelled data is not always available for researchers, mainly in the case of specialised fields such as lung cancer detection. If classes are not distributed equally, the learning process of a machine learning model might get distorted and pay more attention to the popular chunks of data. In this study, we solved these issues by adding new data to the original dataset using data augmentation, which let us expand the range of the data and make it easier for the model to generalise.

Techniques such as data augmentation made it possible to enlarge the dataset by repeating previous rows. Thirdly, when the classes were unbalanced, with very few benign images compared to malignant and normal ones, the way this was handled went beyond data augmentation. When imbalanced data is fed to a classifier, it often ends up providing prioritized answers in favour of the class found most often. In order to deal with this, the data with these class labels was oversampled. To keep the groups of data in balance during training, oversampling means creating additional samples of the less common class. This was done, so that samples from the benign category were enhanced even further, in order to help ensure that the benign number was close to the malignant and normal categories, and to help balance and stabilise the learning procedure.

### *D. Evaluation metrics*

Three popular ways to evaluate a classification model are generally called standard metrics. accuracy, sensitivity, and specificity. The measures are worked out using the model's TP, TN, FP, and FN values.

Accuracy means how often the model gets it right when compared to what actually happened: "(TP + TN) / (TP + TN + FP + FN)" means the accuracy.

It shows what percentage of the predictions, both right and wrong, the model got right.

• "Sensitivity, which people also call recall or the true positive rate, shows how well the model can find out which cases are actually positive" TP divided by (TP + FN) is sensitivity. A higher sensitivity means that the model is good at picking out the real positive examples from the data[17]-[18].

• "The true negative rate, which is also called specificity, shows how good the model is at spotting cases when something is not the problem." TN divided by (TN + FP) is called specificity. By correctly getting rid of examples that shouldn't belong, this number shows how good the model is at not making mistakes and picking the wrong examples. When taken as a whole, these metrics provide a comprehensive understanding of how well a classification model distinguishes between different classes[20].

## III. COMPARATIVE RESULTS

This table lists how many output classes each algorithm can use and how accurate, sensitive, and specific they are in classifying lung cancer, as designed by other researchers.

| Researcher | Number of Classes | Accuracy (%) | Sensitivity (%) | Specificity (%) |
|---|---|---|---|---|
| Al-Huseiny et al. [7] | 2 | 94.38 | 95.08 | 93.70 |
| Kareem et al. [8] | 3 | 89.89 | 97.14 | 97.50 |
| Al-Yasriy et al. [10] | 3 | 93.54 | 95.71 | 95.00 |
| A.A. Abe [19] | 3 | 95.43 | 93.40 | 97.09 |

### *A. Column Descriptions:*

Researcher: The creators of the model are the person or persons in charge. Number of Classes: Normally, the model learns to distinguish between Normal and abnormal using just two classes, like healthy and malignant. Three groups typically refer to normal, benign, and malignant tumors.

Accuracy (%): The share of accurate forecasts among all made forecasts.

Percentage of real cancer cases the model classifies as positive.

Specificity shows how accurately the model spots what is not cancer.

## IV. CONCLUSION

In conclusion, “a really important step forward in medical testing has happened by using machine learning and CNNs, which let them better find out if someone has lung cancer.” CNNs make it easier and faster to find and treat lung cancer in patients because they’re really good at looking at and understanding the information from medical scans. This development could help a lot of people survive and get better results when they get treated. According to research, CNNs can end up giving really good predictions if the data is handled well and the right type of model is chosen. However, before these models can be fully used in the clinic, they need to be checked with real patients in different settings, and they should be checked by health care professionals. To make sure that these technologies are safe and work well for patients, more work needs to be done.

**References:**

[1]. alyasriy, hamdalla; AL-Huseiny, Muayed (2021), “The IQ-OTHNCCD lung cancer dataset”, Mendeley Data, V2, doi: 10.17632/bhmdr45bh2.2.

[2]. B. S, P. R and A. B, "Lung Cancer Detection using Machine Learning," 2022 Salem, India, 2022, pp. 539-543, doi: 10.1109/ICAAIC53929.2022.9793061.

[3]. R. Sathya, P.Rohini, “Breast Cancer Prediction from Multimodal Datasets Using Deep Learning Techniques”, Journal of Next Generation Technology (ISSN: 2583-021X), vol. 5, no. 2, pp. 113-121. April 2025.

[4]. Mohammad Q. Shatnawi, QusaiAbuein, Romesaa Al-Quraan,Deep learning-based approach to diagnose lung cancer using CT-scan images,Intelligence-Based Medicine, Volume 11,2025,100188,ISSN 2666-5212.

[5]. Javed, R., Abbas, T., Khan, A.H. et al. Deep learning for lungs cancer detection: a review. ArtifIntell Rev 57, 197 (2024).

[6]. Sinjanka, Y., Kaur, V., Musa, U.I. et al. ML-based early detection of lung cancer: an integrated and in-depth analytical framework. DiscovArtifIntell 4, 92 (2024).

[7]. Al-Huseiny, M. S., & Sajit, A. S. (2021). Transfer learning with GoogLeNet for detection of lung cancer. Indonesian Journal of Electrical Engineering and computer science, 22(2), 1078-1086.

[8]. Kareem, Hamdalla F., et al. "Evaluation of SVM performance in the detection of lung cancer in marked CT scan dataset." Indonesian Journal of Electrical Engineering and Computer Science 21.3 (2021): 1731.

[9]. KamtaNath Mishra, Alok Mishra, Soumya Ray, Anjali Kumari, SaadMisbahWaris, Enhancing cancer detection and prevention mechanisms using advanced machine learning approaches, Informatics in Medicine Unlocked,Volume 50,2024,101579,ISSN 2352-9148.

[10]. Al-Yasriy, Hamdalla F., et al. "Diagnosis of lung cancer based on CT scans using CNN." IOP conference series: materials science and engineering. Vol. 928. No. 2. IOP Publishing, 2020.

[11]. X. Zhang, S. Li, B. Zhang, J. Dong, S. Zhao, and X. Liu, "Automatic detection and segmentation of lung nodules in different locations from CT images based on adaptive -hull algorithm and DenseNet convolutional network," International Journal of Imaging Systems and Technology, vol. 31, no. 4, pp. 1882–1893, 2021.

[12]. L. Sun et al., "Attention-embedded complementary-stream CNN for false positive reduction in pulmonary nodule detection," Computers in Biology and Medicine, vol. 133, Jun. 2021, Art. no. 104357.

[13]. S. T. Vemula, M. Sreevani, P. Rajarajeswari, K. Bhargavi, J. M. R. S. Tavares, and S. Alankritha, "Deep Learning Techniques for Lung Cancer Recognition," Engineering, Technology & Applied Science Research, vol. 14, no. 4, pp. 14916–14922, Aug. 2024.

[14]. P. G. Mikhael et al., "Sybil: A Validated Deep Learning Model to Predict Future Lung Cancer Risk From a Single Low-Dose Chest Computed Tomography," Journal of Clinical Oncology, vol. 41, no. 12, pp. 2191–2200, Apr. 2023.

[15]. S. Wankhade and V. S., "A novel hybrid deep learning method for early detection of lung cancer using neural networks," Healthcare Analytics, vol. 3, Nov. 2023, Art. no. 100195.

[16]. V. K. Gugulothu and S. Balaji, "An early prediction and classification of lung nodule diagnosis on CT images based on hybrid deep learning techniques," Multimedia Tools and Applications, May 2023.

[17]. S. Zafar, J. Ahmad, Z. Mubeen, and G. Mumtaz, "Enhanced Lung Cancer Detection and Classification with mRMR-Based Hybrid Deep Learning Model," Journal of Computing & Biomedical Informatics, vol. 7, no. 02, Sep. 2024

[18]. M. M. Musthafa, I. Manimozhi, T. R. Mahesh, and S. Guluwadi, "Optimizing double-layered convolutional neural networks for efficient lung cancer classification through hyperparameter optimization and advanced image pre-processing techniques," BMC Medical Informatics and Decision Making, vol. 24, no. 1, May 2024, Art. no. 142.

[19]. A.A. Abe, M. Nyathi, A.A. Okunade, W. Pilloy, B. Kgole, N. Nyakale,A robust deep learning algorithm for lung cancer detection from computed tomography images,Intelligence-Based Medicine,Volume 11,2025,100203,ISSN 2666-5212.

[20]. G. Srija, E. Bhargavi, G. Sai Teja, B. Kotesh, P. Anusha, B. Santhosh Kumar, “Synthetic Data Generation and Machine Learning for Enhanced Heart Disease Risk Prediction”, Journal of Next Generation Technology (ISSN: 2583-021X), 5(5), pp. 22-30. July 2025.