Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

Lung Cancer Detection using Machine Learning

Anurag Tiwari[1], Ms. Mariyam Kidwai [2] Assistant Professor , Ms. Ambreen Anees [3] Assistant Professor

[1] Research Scholar Dept. Of Computer Science and Engineering, Integral University, Lucknow, U.P.,

[2] Dept. Of Computer Science and Engineering, Integral University, Lucknow, U.P.,

[3] Dept. Of Computer Science and Engineering, Integral University, Lucknow, U.P.

Abstract - It is most fatal in the world and this combined with complete failure to give proper early detention sites has made it the most dominant cause of cancer worldwide. When detected in the early stages, lung cancer has great possibilities of successful treatment. Machine Learning (ML) and Artificial Intelligence (AI) technologies have been on the increase in recent years becoming a useful mode of addressing a range of problems related to healthcare such as diagnostic accuracy. This study aids the understanding of Machine Learning practice in relation to multiple classifications in Lung Cancer detection via a large biomedical imaging data by analysis as well as patient-specific clinical characteristics. We got to learn the various Machine Learning and Deep Learning algorithms, such as the Logistic Regression, Decision tree, random forest, Support Vector machine (SVM), convolutional Neural Network (CNN) and their classification were inquired. Publicly available datasets such as LIDC-IDRI dataset; Kaggle Lung Cancer dataset were used for training and testing our models in the study. The methodology used was: Data processing→ Feature extraction→Feature selection→Model training and performance evaluation that the performance can be quantified in terms of quantitative metrics like accuracy, precision, recall, and F1-score and ROC-AUC score. The obtained outcomes of the experiments reveal that the proposed methods of deep learning, specifically, the application of CNN, are the most precise and can be scaled up in terms of counting the lung nodules, which turn out to be cancerous in accordance with the images of the CT scan. The researchers add that AI-assisted diagnostic systems mean the minimization of human error, decreasing of stage detection, and helping radiologists to make clinical decisions. Future health care and machine learning based lung cancer detection systems have wide usage but reliance on data set, excessive complexity of computing etc are drawbacks related to it.

Keywords - Lung Cancer, Machine Learning, Artificial Intelligence, Deep Learning, Convolutional Neural Network, Support Vector Machine.

1. INTRODUCTION

1.1 Background of Lung Cancer

Personally, lung cancer is one of the deadliest diseases present in millions patients throughout the world. “It happens when the cells in your lungs multiply and begin to increase, forming larger lumps restricting normal breathing. Lung cancer is divided into two classes, non-small cell lung carcinoma (NSCLC) and small cell lung carcinoma (SCLC). For these cases, the largest subgroup was defined by diagnosed NSCLC. The main cause of lung cancer is tobacco smoking but in addition to the tobacco smoke itself, air pollution sources, occupational exposure to toxic chemicals and genetic susceptibility play a role in the development of the lung cancers [1].

In global health, lung cancer represents one of the top five causes of cancer-related death thanks to its chemotherapeutic responsiveness and late stage diagnosis. In particular, lung cancer incidence in India is rising steadily for the last 10 years owing to prolonged tobacco smoking, increasing urban pollution & lifestyle changes [2]. While this condition is usually experienced in men, there are increasing cases of women with the disease too. These are behaviours and allergy cough symptoms, which, at first stage, often are dismissed or misdiagnosed by the medical field: persisting coughing, chest pain associated with unintentional weight loss and breathing problems. Consequently, the polishing of diagnostic methods and allowing the detection at an early age has taken centre stage in the ambitions of contemporary health care systems.

1.2 Need for Early Detection

Lung cancer is a major cause of death in most countries hence early diagnosis and treatment is of high essence in terms of survival as well as reduction of death among patients. Surgery, chemotherapy, radiation therapy and targeted therapies are more effective if lung cancer is caught early.

The conventional diagnostic methods tend to identify the disease though at a later stage.

Traditional approaches like biopsy or chest X-rays and cytology of sputum or interpretation of CT scans by radiologists are time consuming and associated with human

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

factors, culminating in errors in diagnosis [3]. The absence of advanced healthcare facilities and even specialists in most developing countries adds another layer of complexity to it foe early diagnosis. Unequivocal and timely recognition not only incurs substantial expenses of treatment but adversely affects the possibility of recovery. Therefore, a compelling need exists to develop intelligent automated systems to aid the healthcare professionals in accurately identifying cancerous pattern efficiently. The recent advancements in medical imaging and computational sciences have opened up opportunities for making automated diagnostic models which can potentially aid the quicker, robust lung cancer detection.

1.3 Machine Learning in Health Care

AI and ML have paved a transformational path in the healthcare system for seamless clinical decision-making through intelligent analysis of a large corpus of medical data. What Machine Learning means: it is a sub-field of AI that allows computer systems to infer trends from previous data, and predicts with minimal human effort. In health care, "ML methods can be applied in many areas from disease diagnosis, patient monitoring, medical image analysis, drug discovery and predictive analytics [4]. For example, ML algorithms have been shown to perform image classification algorithms on CT scan images for lung cancer detection and accurately classify tumor type and likelihood of cancer by identifying the most abnormal nodules. Support vector machine (SVM), decision trees, random forest, naïve Bayes, logistic regression and Convolutional neural networks (CNN) are some of the algorithms that could be used in medical diagnosis. They reduce human errors, make the diagnostic process faster and assist in better clinical decision-making. The deep learning models, especially Convolutional Neural Networks or CNNs are used to achieve great performance on image based cancer detection due to their inherent ability for automatic extraction of important features from medical images [5]. By being implemented in health systems, AI and ML can be used to increase diagnostic accuracy, lighten radiologist workload and offer wallet-friendly options for early cancer screening.

1.4 Objectives of the Study

The main goal of this paper is to perform a comparison study of Machine Learning algorithms for lung cancer detection since the early stages using comprehensive data and medical imaging techniques. The study assesses various ML algorithms and compares the performance, accuracy, precision, recall, and efficiency. Another purpose is knowing how AI-based systems can help healthcare professions in making quicker and more accurate diagnosis. In addition, the study aims to find the best-fit machine-learning model for predicting and classifying early-stage lung cancer. In addition, the study aims to investigate the potential of CT scan image processing in enhancing diagnosis and reducing mortality that is associated with failing to detect positive results on time. This paper seeks to bolster the growing area of computer-aided diagnosis by analyzing past studies and existing technologies.

1.5 Scope of the Study

This study only focuses on the use of Machine Learning techniques in detecting and predicting lung cancer. This research is concerned with image-based diagnostic systems using CT scan datasets related to publicly available medical data. We study wide range of ML/DL algorithms to evaluate the right methods for identifying cancerous lesions in lung tissues. Performance metrics are then compared on various models conducted in the study namely, Accuracy, Sensitivity, Specificity and F1-score. But this was neither a clinical trial in real time, nor with patients. This research is largely academic and theoretical, aimed at assessing the prospects of deployment of AI-enabled diagnostic systems in healthcare organisations. Finally, the end of the study describes the impact of advancements in hybrid AI models, real-time screening systems and cloud-based healthcare applications.

2. LITERATURE REVIEW

2.1 Traditional Detection Techniques

INTRODUCTION The usual practice of Lung cancer diagnostics includes chest X-ray, sputum cytology, bronchoscopy biopsy and CT methods. Appropriate members of the medical community have used these techniques for many years. One of the first tests that you will execute is a chestx–raytoassessinitiallyhovertiveastiorelcalltheseverysmallnodulesortumorscerteo (1–5) [6]. High-resolution CT also offers a better identification of pulmonary defects with enhanced sensitivity. Biopsy is the gold-standard for cancer diagnosis but it is invasive, cumbersome, costly and time-consuming. Medical images interpretation also requires a lot of skills, which leads to inter-observer variability and more susceptibility to human errors. These issues also prompted researchers to investigate computational methods for enhancing diagnostic accuracy and clinical decision support [7].

2.2 Machine Learning-Based Detection

Machine learning based detection systems have been identified among the most promising methods to enhance

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

productivity towards lung cancer detection. Prediction Models using SVM, Decision Tree, Random Forest, KNN and Logistic Regression Model In this work the researchers invented prediction model. They feed the data on either patients or medical images, and classify them as benign–tumors or malignant tumors.

Several previous studies concluded that ML approaches had the potential to improve diagnostic accuracy at a lower computational cost [8]. Example: You are trained with data upto 2023October SVM for example was very popular and was used in machine learning for many classification problems in particular the medical-data due to high dimensional input. While Random Forest algorithms aggregate multiple decision trees providing more robust results, Logistic Regression is used for probability-based predictions [9]. Such models are helpers that automate these analyses and rely less on human interpretation of the images, making a more resource-efficient use of health care resources. They also applied different feature extraction methods from Machine Learning to achieve better efficiency in identifying Cancerous nodules found on CT scan images.

2.3 Deep Learning Approaches

At present time Deep Learning has turned into one of the models employed by in reality image analysis [4] and (/or) lung cancer identification[5.6]. Is an because CNNs can comprehend pertinent features of images without established features, also referred to as. Various models based on convolutional neural network (CNN) have achieved high accuracy in recognition of lung nodule, tumour classification and detection of even primitive stages of cancer with CT scan images [10]. Deep learning requires large datasets and thus keeps on using transfer learning and hybrid deep learning model for performance enhancement. This method is not much dependent on features engineering can show improvement in accuracy of diagnosis with the increase of image data. When it comes to image-based classification tasks, one of the scenarios where CNN models beat such traditional machine learning algorithms. Deep learning systems perform better on larger datasets with plentiful computational resources and substantial training times [11]. Although these challenges are present, still deep learning is one of the main components for advanced healthcare intelligent systems development focused on automatic cancer diagnostic.

2.4 Research Gap

Although much has been accomplished regarding lung cancer detection employing Machine Learning and Deep Learning technologies, several research gaps still exist. Nevertheless, in spite of many current works towards accurate modeling, the literature largely ignores further essential aspects for machine learning application including model interpretability, computational cost and true clinical implement ability. Some models are trained exclusively on a single data set, potentially being ill-suited to withstanding the variability which is inherent in other patient populations. Furthermore, obtaining high-quality annotated medical datasets still continues to be a problem for researchers. And in many cases experiments are performed without directly comparing various algorithms under the same experimental conditions. Also, deployment of hospital infrastructure and clinical workflows by AI-based systems is still relatively limited due to privacy issues and technical complexity in addition to ethics. Simultaneously, it also underscores the necessity for more and better utilized models which are not only capable enough yet name-profitability-saving-sustainable-scavenger solutions in healthcare practice[2].

3. METHODOLOGY

This work follows the method of a machine learning based system for discovering lung cancer by using data sets which includes medical images. The methodology is composed of steps such as soaking a data set, preprocessing (cleaning and preparing the data for analysis), feature extraction, model training/test and analysing the results [12].

This technique is primarily for developing a intelligent diagnosing system to classify CT scan images and aims to provide those alongside a patient as cancerous nodules. In this paper, different machine learning and deep learning algorithms are implemented and compared to optimize the best model for lung cancer classification. A systematic workflow has been followed in the study which starts with data acquisition and ends with predictive analytics and assessment of model performance.

The nature of overall research design is experimental and based on supervised machine learning. In supervised learning models, there are labeled datasets to train the model with pre-defined output categories (e.g. cancerous and non-cancerous cases). It involves the use of image processing alongside classification algorithms in order to predict lung cancer with more accuracy. This workflow also allows for visualization and comparison of various models on performance metrics e.g. accuracy, precision, recall, f1-score and sensivity.

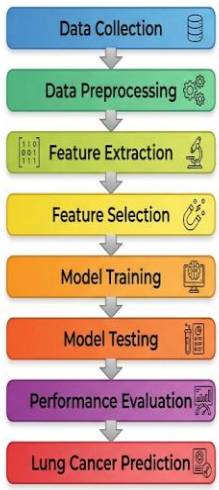
3.1 Proposed System Architecture

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

The basic figure of the proposed methodology is as shown below The system architecture of the modeled lung cancer detection scheme containing multiple connected stages functioning for automated detection through various methods in an efficient manner. In a first step, public medical repositories are used to collect CT scan images and patient datasets. These datasets pass a preprocessing steps which include image resizing, normalization and noise removal & segmentation to improve the quality of the images and remove unnecessary information. Once the input is preprocessed, feature extraction is carried out to select relevant features like texture, shape, size and intensity of lung nodules.

The extracted features are then passed to different ML algorithms such as Logistic Regression, Random Forest, Support Vector Machine (SVM), Decision Tree, Convolutional Neural Network (CNN). Essencially, each model is trained on training datasets and validated on testing datasets. The models identify the output as either normal or suspect lung tissue. Then performance metrics are computed to evaluate how well the models have performed and ascertain the winning algorithm.

Figure: Proposed Methodology Flowchart



3.2 Dataset Description

Machine Learning models are heavily dependent on Dataset used for both Training and Testing. The datasets consist of publicly available lung cancer data and were used to validate that all experiments could yield the same result. In this paper, we focus on the following key datasets: LIDC-IDRI dataset and Kaggle lung cancer dataset.

(a) LIDC-IDRI (Lung Image Database Consortium and Image Database Resource Initiative) dataset: A database of thousands of CT scan images of the thorax with lesions annotated by professional radiologists. This dataset is widely used with regards to lung cancer detection work since it has a lot of contextual detail in both the studies (ideal for imaging processing and classification workloads).

Datasets The Kaggle Lung Cancer dataset contains various features such as age, smoker or not, anxiety level scale, fatigue scale, chest pain scale along with the frequency of cough and duration of cough that may show lung cancer signs. Dataset for predicting clinical features and patient records. Both datasets are divided into training & testing sets so that we can systematically compare the performance of our models.

3.3 Data Preprocessing

Data preprocessing is one of the quintessential stages of machine learning because raw medical data usually contains noise, inconsistencies and missing values that harms model performance. The study covers multiple preprocessing steps to gain better quality datasets in order to utilize it in a way that gives more efficient performance against the model.

Image datasets preprocessing includes resizing images, converting from RGB to gray/saturation normalization, image filtering and image segmentation. Uniformity dictates that all CT scans will be resized to the same shape, along with normalising pixel intensity values as a separate step. Gaussian Filtering and median filtering are noise removal filter which we will be implemented on clean images. Segmentation Examples The techniques which separate regions of the lung and masses of abnormal nodules on the surrounding tissues.

Encoding categorical variables into numerical values and methods of data imputation to cope with missing data were the best applications to tabular datasets. Meanwhile, ensure that all features have the same contribution in the scaling aspect before being inputted into the model to train. This changing to thee Makes Greater Accuracy with Less Daintly.

3.4 Feature Selection

In this respect, feature selection is the technique of selecting feature (or attributes) of the data set that attain the greatest information gain in the prediction of lung cancer. The identification of features is a very important feature in order to prevent overfitting and enhance the performance of the model and to decrease the amount of dimensions.

The important imaging-based features of the CT scanned images of the nodules used in the analysis process are broken

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

down: Texture, shape, edge vs strength/nodule size. Patient datasets included smoking history, having chronic cough, chest pains of age and difficulty in breathing. Correlative and statistical analysis reveals the importance of features of lung cancer. Feature Selection, some methods to help the models reach a better performance are principal component analysis (PCA), correlations analysis and recursive feature elimination.

This way you can decrease the computation massively while maintaining high diagnostic accuracy using only the most informative features.

3.5 Machine Learning Algorithms Used

This paper implements various machine learning and deep learning algorithms to determine their performance in lung cancer diagnosis.

Logistic Regression

Logistic Regression is a supervised classification algorithm, that is applied for binary prediction problems. It is predict Lung Cancer Occurrence by giving features. The algorithm is easy to compute and applies to a medical classification problem in smaller datasets.

Random Forest

Random Forest is an ensemble learning algorithm that combines several decision trees to improve the accuracy of predictions and control for overfitting. It performs well on huge datasets and complex categorization tasks. The algorithm performs quite robustly and provides sound feature importance analysis.

Support Vector Machine (SVM)

Support Vector machine is one of the most common classification algorithms in the area medical image analysis. SVM: SVM identifies the best hyperplane that has maximum margin separating cancerous and non-cancerous cases. It has been shown to work well in high-dimensional data sets and yields a classifier with high accuracy.

Decision Tree

Decision Tree is a tree-like structured classification algorithm which splits the data into competing branches based on features and their values. It is a transparent model and thus easy to interpret, which is beneficial for healthcare related applications. On the other hand, if optimized inappropriately, it may lead to overfitting

Convolutional Neural Network (CNN)

CNN: CNN is a specialized deep learning architecture for processing images and pattern recognition problems. It performs automated feature extraction from CT scan images by means of convolutional layers, pooling layers and fully connected layers. Because Convolution Neural Networks (CNN) having the capability of learning complicated features from images without intervention by humans, CNN models are very efficient for detecting lung nodules and image-based classification tasks.

3.6 Training and Testing Process

The sampled datasets are later separated into two parts as training data and testing data.

Commonly, it should take around 70% -80% of the data for training the model and rest are used in testing or validation. Machine learning models learn the patterns and relationships from the input data at training. Lastly, these models are validated on out-of-sample testing data in order to present how well they can predict.

Additional cross-validation methods are also used to ensure model stability and prevent over-fitting. To run hyperparameter tuning of the model via grid search and even use certain optimization approaches to make it more efficient. The output of both the models are compared in order to identify the most suitable algorithm in detecting lung cancer.

3.7 Evaluation Metrics

This study makes a comparative 0 of different algorithms using various performance measures.

The key evaluation metrics are:

Accuracy: The fraction of correctly classified instances

Precision It informs on the number of positive cases among the number of predicted positive cases which were truly positive.

Recall (Sensitivity): Shows the models ability to predict a cancer case correctly.

F1-Score: Harmonic Mean of Precision and Recall

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

Specificity: Measures the ability to correctly identify a non-cancer case.

Confusion Matrix: Shows true positive, false-positive, true negative and false-negative.

ROC Curve and AUC: Assess classification performance across various threshold values, in terms of ROC curve and area under the curve (AUC).

Such metrics currently help investigate an explicable and clinically applicable lung cancer detection system.

4.DATASET AND TOOLS USED

In this study, publicly accessible datasets and state of the art computational tools are employed for classification through Machine Learning models in lung cancer detection. The datasets are LIDC-IDRI dataset, Kaggle Lung Cancer Dataset etc. LIDC-IDRI: A dataset of images and corresponding annotations provided by a panel of radiologists on CT scan images which could be useful for image based cancer detection. Details on symptomatic and clinical data of patients that can help in predictive analysis and classification problems using the Kaggle dataset.

Throughout the implementation and experimentation, different software tools and programming technologies are applied. Python: we used because its highly simple, flexible and also has the most machine learning libraries [13]. We will use Jupyter Notebook as development environment visualising, coding and performing the model. implementation of basic ML algorithms using scikit-learn: Logistic Regression, SVM, Random Forest and Decision Tree. Tensorflow and Keras Frameworks used for deep learning models such as CNN. OpenCV is responsible for image preprocessing and computer vision operations, NumPy and Pandas are used for numerical computation and data manipulation.

Hardware Requirements The hardware required for this study are the computer system with minimum Intel Core i5/i7 processor, 8 GB RAM and adequate storage capacity. You can take advantage of GPU support for faster deep learning model training. Software requirements You need Windows OS/Linux, python environment, Jupyter Notebook to run this project and install the following modules: TensorFlow (2.0), OpenCV, Scikit-Learn, NumPy, etc. All these tools together give a fast way for training and testing lung cancer identification models using Machine Learning.

5.RESULTS AND ANALYSIS

This section describes the performance analysis of various ML & DL algorithms used for lung cancer detection. The raw data were the CT scan images and clinical patient data, out of which you can be able to derive variables. To carry out this analysis we will search the accuracy of prediction using each of the algorithms and determine the best way to efficient ordinal high dimensional representation that can distinguish best lung cancer. Evaluation of the model was conducted with a standard set of performance metrics such as accuracy, recall F1 score confusion matrix and ROC & area under curve (AUC). The experimental findings give hope on the use of Machine Learning methods in diagnosing lung cancer during its initial stages, which will support the health practitioners with features that assist in making a clinical decision.

The following algorithms were the ones that were employed in this research: Logistic Regression Decision Tree Random Forest Support Vector Machine (SVM) Convolutional Neural Network (CNN) Each of the following algorithms were executed under uniform conditions based on train and testing sets. This could be due to difference in prediction accuracies and classification performance of models Overall, deep learning approaches, particularly CNNs performed the best among these as those methods were most successful in automatically extracting features from CT scan data. However, with the appropriate preprocessing and feature selection techniques many traditional machine learning models can also produce quite good results.

5.1 Experimental Results

Experimental analysis was carried out on open access datasets which also contain publicly accessible LIDC-IDRI and Kaggle Lung Cancer datasets. To be able to evaluate our models effectively, we did a random extraction of training and testing data of each of the household types at each NZD level, 80:20. You programmed yourself to conduct a detailed data analysis to check the effect of model parameters on the time-to-market performance using some preprocessing strategies (normalizing, segmentation, and feature extraction). Classifier accuracy, precision, recall and F1-score are computed to find the performance of every algorithm. Accuracy is the number of true positives + true negatives out of all test cases, while precision and recall show how reliable our positive predictions are (precision) as well as How Many Cancer Cases Actually Were Found by Our Algorithm (recall). The F1 score or F-score is a weighted harmonic mean of Precision-Recall and decides the fake positives and false negatives classification report

CNN shows better accuracy in comparison with all other algorithms due to its high capability of feature learning as

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

revealed by the experimental results. High classification performance and confidence scores were returned by both Random Forest and SVM. While Logistic Regression and Decision Tree have a much lower precision, they still score satisfactory results for basic classification problems.

Table 5.1 Performance Evaluation of Machine Learning Models

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	87.2	85.4	84.8	85.1
Decision Tree	89.1	88.2	87.6	87.9
Random Forest	93.5	92.8	93.1	92.9
Support Vector Machine (SVM)	94.2	93.7	93.9	93.8
Convolutional Neural Network (CNN)	97.1	96.8	96.5	96.6

The results were evident and CNN outperformed all other algorithms with the highest accuracy of 97.1%. Moreover, the precision and recall values demonstrate that CNN managed to successfully identify both cancerous/non-cancerous cases, with not much misclassification. Likewise SVM and Random Forest were able to achieve nice accuracy results classification as efficiently remain good techniques by using effectively also which creates them users recurse in good possible effective of lung cancer prediction.

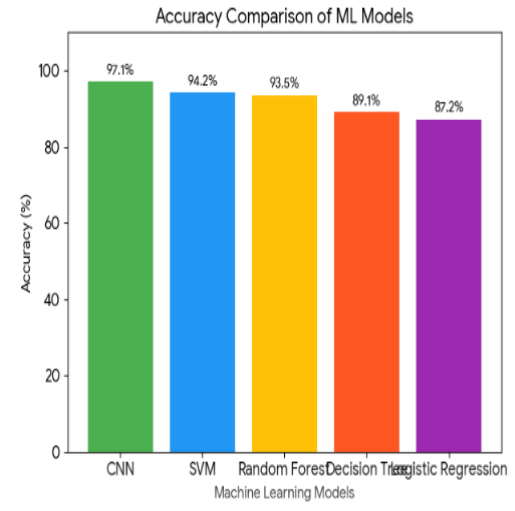


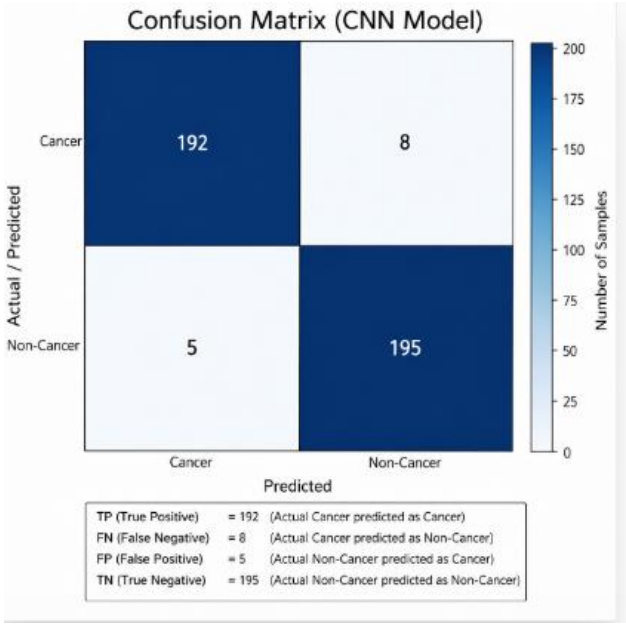
Figure2: Accuracy Comparison Graph

Confusion Matrix Analysis

Confusion Matrix: Confusion matrix is a relevant tool because that provides an insight about model predictionTPTNTFPFNPThen more the T.P and T.N _perdictition, better is the model performance.

Table 5.2 Confusion Matrix for CNN Model

Actual / Predicted	Cancer	Non-Cancer
Cancer	192	8
Non-Cancer	5	195



The CNN model placed in National Cancer Grid Program correctly classified a huge number of cancerous and non-cancerous cases with minimal misclassification, as shown by the confusion matrix. This shows the strength of deep-learning techniques in medical imaging and automatic detection of cancer.

ROC Curve Analysis

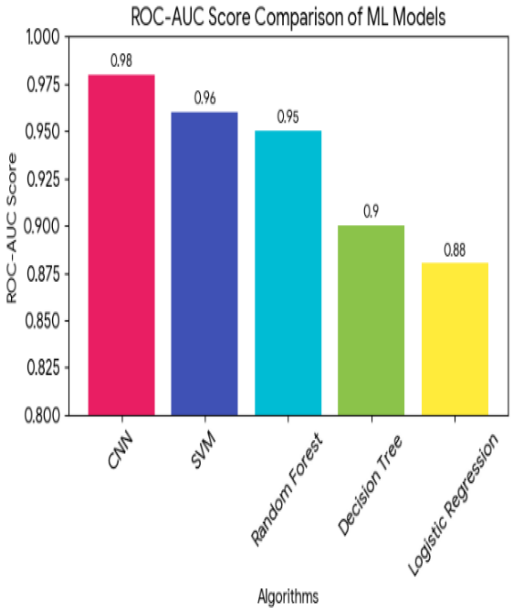
Our choice of measurement of the classification performance of Machine Learning models is the Receiver Operating Characteristic operation (ROC) curve. He is an important plot that relates to the sensitivity and specificity of different threshold levels. AUC closer to 1 indicates a better classification.

Table 5.3 ROC-AUC Comparison of Algorithms

Algorithm	ROC-AUC Score
Logistic Regression	0.88
Decision Tree	0.90
Random Forest	0.95

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

SVM	0.96
CNN	0.98



ROC-AUC results show that for AUC, CNN obtained the highest value 0.98 indicating that overall this model could classify very well and is appropriate to be used as a reliable diagnostic test.

5.2 Performance Comparison

Following is a comparative study of various Machine Learning methods showing high disparities in prediction efficiency and classification accuracy. It performed well with binary classification problems but could not classify complex image patterns in case of Logistic Regression. Decision Tree achieved better results and was more interpretable than Logistic Regression, but had a higher tendency to overfit the data.

Random Forest showed improved stability and accuracy because as it averages predictions over multiple decision trees to make sure the overall prediction is reliable. SVM also showed reasonably good results due to its ability to handle high-dimensional datasets and modeling classification boundaries. The highest performance was for the CNN and as it can automate feature extraction of the image data while able to learn intricate patterns from CT scan images.

The comparison also shows that deep learning approaches are more suited to image based medical diagnosis while traditional machine learning methods can perform this task with images as well but lack in feature extraction. Nevertheless, traditional algorithms do use minimal computational resources and are also the more reliable option when working with smaller datasets or in symptom based prediction systems.

Table 5.4 Comparative Analysis of Algorithms

Parameter	**Logistic Regression**	**Decision Tree**	**Random Forest**	**SVM**	**CNN**
Classification Accuracy	Moderate	Moderate	High	High	Very High
Training Time	Low	Low	Medium	Medium	High
Feature Extraction	Manual	Manual	Manual	Manual	Automatic
Computational Complexity	Low	Low	Medium	Medium	High
Suitability for Image Analysis	Limited	Limited	Moderate	Good	Excellent

This table indicates that CNN offers better performance for image-based lung cancer detection; whereas, Random Forest and SVM provide viable alternatives with lower computational requirements.

5.3 Discussion of Findings

Occupying this vacuum, it is possible to obtain more accurate detectives to identify lung cancer and extract the necessary information that will help in the early diagnosis process with more Machine Learning and Deep Learning techniques. Automated classification systems leads to lower amount of manual scan interpretation and has therefore led to far less opportunity for diagnosis errors in verification of test results The result produced confirms the premise that CNN-based deep learning models perform very well at deliberately detecting the lung nodules and pathological (cancerous) growths from CT scan images [14]. This study also demonstrates that feature extraction and pre-processing are essential to enhance model performance. Better prediction rates have been achieved for all algorithms due to proper normalization, segmentation and feature selection [15]. The best results are achieved using CNN but these require a huge dataset and take long to train, which can make this option unrealistic for an often-limited computational resource.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

SVM, Random Forest, on the other hand have provided competitive performance but with low computational complexity required for real life applicability in health care systems with limited resources [16].

In summation, the complete analysis suggests that Machine Learning system utilization in institutions can increase efficiency and streamline early-stage cancer detection via aiding surgeons & radiologists within healthcare thereby enhancing outcome at a higher level [17]. Nevertheless, large clinical datasets with translatable and implementable real-time use may enhance the performance of AI-based systems in variable healthcare contexts by validating its diagnostic accuracy.

6.ADVANTAGES AND LIMITATIONS

Machine Learning based lung cancer detection systems have many merits in existing healthcare and medical diagnosis. A great benefit is something like the faster diagnosis. The traditional method of diagnosis involves lengthy image analysis and interpretation by radiologists, whereas machine learning (ML) algorithms have the ability to analyze massive volumes of medical data in a short period [18]. It allows clinicians to make quicker clinical decisions and succeed in initiating treatment sooner. One such advantage that should not be overlooked is diagnostic accuracy. Deep learning algorithms (and CNNs in particular) aid in identifying complex features and abnormalities that are contained in CT scan data that would not necessarily be spotted by human eyes (when observed manually). This leads to increased likelihood of correct detection of cancer.

The human error is also minimised in a medical diagnosis implemented using ML systems. Just like in fatigue, work loads and differences in quality of pension by radiologists are also false positives or negatives as in case with pathologist specialists CT scan picture interpretation. Trustworthy diagnosis will be formed with such an automated system based on AI, as you will be able to conduct a standardized and objective point of view. The early diagnosis of lung cancer is another important aspect of these systems [19]. Detection of cancerous nodules at an early age enhances the chances of successful cure and survival by the patients. This is paramount to any presentational healthcare programs, especially those that are screening patients of cancer since ml models begin to understand when we are not normal. However, there are many limitations for machine learning based lung cancer detection systems. The single most important limitation is that it relies on the quality of dataset. Actually, how good your ML could be can heavily depend on the dataset itself as they say images are one of most important factors in the kingdom of well-labeled. It can lead to incorrect forecasts or skewed outcomes in return for working with inferior quality datasets.

Deep learning models also demand a lot of computation. For the efficient end-to-end training and processing of algorithms like CNN, powerful hardware systems with substantial memory capacity as well as GPU support are required for quick access to latent patterns in data, which unfortunately not every healthcare institution can facilitate.

It is also worth mentioning that it pours you a high risk of overfitting. This is because sometimes, well-behaved models have a tendency to learn accurately from training data but when it comes to external testing on actual real-world problems their generalization capabilities are not that great. This way, the system becomes less reliable in real-world mechanisms of healthcare institutions. In addition, there is still inadequate real-world clinical validation for many Machine Learning models. Most studies are based on existing real data sets in experimental settings that allow for relative control, but the actual hospital environment has significantly higher complexity and variability. As a result, these systems require additional extensive clinical testing and validation before being fully incorporated into routine medical practice.

7.FUTURE SCOPE

Future Scope of Machine Learning for Lung Cancer Awareness of Future of Machine learning frontier- - Detection- Soon AI and medical imaging will skyrocket, and become massive. A future perspective of interest is an improvement by deep learning. This facilitates the learning the most complex neural networks architectures in order to have better features extraction, decrease false prediction hits and gaining higher overall diagnosis accuracy [20]. Another option is to engage more sophisticated methods and techniques that can help the lung cancer detection systems be more reliant and interpretable, e.g., transfer learning, hybrid deep learning models, explainable AI. A second direction much needed in future is development of real-time detection systems which can process CT scan images as soon as they are available for supporting patient diagnosis. Such systems could help providers in the field or reduce time spent on clinical decision making. Furthermore, Machine Learning along with Internet of Things (IoT) and cloud based Healthcare systems could enable remote monitoring with data shared for centralized analysis in medicine [21]. It will be mainly beneficial in the remote and unserved regions, where access to specialist achieved centers is scarce.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

According to reports, AI-driven radiology will in fact be an integrated divider inside the future state of social insurance systems. So in this aspect, Machine Learning models are basically a smart aid system which helps radiologists by pointing out abnormalities in regions of suspected nature and automatic diagnosis suggestions in images. Moreover, more data and a higher variety of data input; helps refine the learnings for the model (generalisation), together with minimised false positives. You can also take the Hybrid Models (Traditional ML + Deep Learning) approach for performance gain and less computation. When accomplished, it has the potential to revolutionize lung cancer diagnosis into a fast, precise and incredibly low-priced healthcare device.

8.CONCLUSION

Bottom line Machine Learning has become an effective, high speed and high influence technology in the process of detecting and diagnosing lung cancer. Lung cancer has been on the rise globally, and early detect methods can optimize patient outcomes. Most diagnostic techniques are hampered by delayed detection, need for expert assessment and vulnerability to errors while interpreting the results. Thus, the introduction of Machine Learning and Deep Learning methods in health care systems is a good remedy to overcome these problems by allowing fast, easy and accurrate analysis on medical data and CT scan images.

This study shows that Machine Learning algorithms have significant potential in Ph1G whose implementation can facilitate effective and accurate lung cancer prediction, but it also depends on practitioners as well in making clinical decision After comparison of some algorithms, CNN module was selected over other algorithm due to better performance by achieving higher precision and recall score along with model classification. Support vector machines (SVM), random forest also gave great results and provide a simple to implement solution targeting classification problems in Medical datasets.

It has many practical implications in the modern health care. Lung cancer detection systems that are driven by AI reduce the workload of diagnostic work and support radiologists in reading images, help to identify those areas where there is a human error possibility due to fatigue or wrong normalisation, and assist with identifying the specific areas visually affected by cancer at an early stage. Finally, a timely diagnosis improves treatment efficacy and subsequently decreases the burden of mortality associated with lung cancer". However these systems do have limitations such as dependency on dataset and increased complexity, though with the rapid advancements of Artificial Intelligence along with medical imaging technology will lead to improvements in their performance.

Hence it is pretty much plausible to believe that Machine Learning based lung cancer detection can take a central stage in the future healthcare and medical diagnostic fields.

REFERENCE

[1] Radhika PR, Nair RA, Veena G. A comparative study of lung cancer detection using machine learning algorithms. In2019 IEEE international conference on electrical, computer and communication technologies (ICECCT) 2019 Feb 20 (pp. 1-4). IEEE.

[2] Rahane W, Dalvi H, Magar Y, Kalane A, Jondhale S. Lung cancer detection using image processing and machine learning healthcare. In2018 International conference on current trends towards converging technologies (ICCTCT) 2018 Mar 1 (pp. 1-5). IEEE.

[3] Bharathy S, Pavithra R. Lung cancer detection using machine learning. In2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC) 2022 May 9 (pp. 539-543). IEEE.

[4] Joshua ES, Chakkravarthy M, Bhattacharyya D. An Extensive Review on Lung Cancer Detection Using Machine Learning Techniques: A Systematic Study. Revue d'Intelligence Artificielle. 2020 Jun 1;34(3).

[5] Asuntha A, Srinivasan A. Deep learning for lung Cancer detection and classification. Multimedia Tools and Applications. 2020 Mar;79(11):7731-62.

[6] Raut S, Patil S, Shelke G. Lung cancer detection using machine learning approach. International Journal of Advance Scientific Research and Engineering Trends (IJASRET). 2021 Jan.

[7] Elnakib A, Amer HM, Abou-Chadi FE. Early lung cancer detection using deep learning optimization.

[8] Tekade R, Rajeswari K. Lung cancer detection and classification using deep learning. In2018 fourth international conference on computing communication control and automation (ICCUBEA) 2018 Aug 16 (pp. 1-5). IEEE.

[9] Pawar VJ, Kharat KD, Pardeshi SR, Pathak PD. Lung cancer detection system using image processing and machine learning techniques. Cancer. 2020;3(4):2020.

[10] Pradhan K, Chawla P. Medical Internet of things using machine learning algorithms for lung cancer detection. Journal of Management Analytics. 2020 Oct 1;7(4):591-623.

[11] Wahab Sait AR. Lung cancer detection model using deep learning technique". Applied sciences. 2023 Nov 20;13(22):12510.

[12] Wu Q, Zhao W. Small-cell lung cancer detection using a supervised machine learning algorithm. In2017 international symposium on computer science and intelligent controls (ISCSIC) 2017 Oct 20 (pp. 88-91). IEEE.

[13] Kalaivani N, Manimaran N, Sophia DS, D Devi D. Deep learning based lung cancer detection and classification. InIOP conference series: materials science and engineering 2020 Dec 1 (Vol. 994, No. 1, p. 012026). IOP Publishing.

[14] Rehman A, Kashif M, Abunadi I, Ayesha N. Lung cancer detection and classification from chest CT scans using machine learning techniques. In2021 1st international conference on artificial intelligence and data analytics (CAIDA) 2021 Apr 6 (pp. 101-104). IEEE.

[15] Abd Al-Ameer AA, Hussien GA, Al Ameri HA. Lung cancer detection using image processing and deep learning. Indones. J. Electr. Eng. Comput. Sci. 2022 Nov;28(2):987-93.

[16] Gayap HT, Akhloufi MA. Deep machine learning for medical diagnosis, application to lung cancer detection: a review. BioMedInformatics. 2024 Jan 18;4(1):236-84.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

[17] Dutta AK. Detecting Lung Cancer Using Machine Learning Techniques. Intelligent Automation & Soft Computing. 2022 Feb 1;31(2).

[18] Yamini B, Sudha K, Nalini M, Kavitha G, Subramanian RS, Sugumar R. Predictive modelling for lung cancer detection using machine learning techniques. In2023 8th International Conference on Communication and Electronics Systems (ICCES) 2023 Jun 1 (pp. 1220-1226). IEEE.

[19] Shakeel PM, Burhanuddin MA, Desa MI. Lung cancer detection from CT image using improved profuse clustering and deep learning instantaneously trained neural networks. Measurement. 2019 Oct 1;145:702-12.

[20] Bhatia S, Sinha Y, Goel L. Lung cancer detection: a deep learning approach. InSoft Computing for Problem Solving: SocProS 2017, Volume 2 2018 Oct 31 (pp. 699-705). Singapore: Springer Singapore.

[21] Meeradevi T, Sasikala S, Murali L, Manikandan N, Ramaswamy K. Lung cancer detection with machine learning classifiers with multi-attribute decision-making system and deep learning model. Scientific Reports. 2025 Mar 12;15(1):8565.