IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134153

# Supervised Machine Learning Approach for Lung Cancer Diagnosis

**Prathima L[1], Rakshitha S C[2], Sanjana R[3], Yuktha Muki V[4]**

Assistant professor, Department of Information Science and Engineering, Jawaharlal Nehru New College of Engineering, Shivamogga, India[1]

Student, Department of Information Science and Engineering, Jawaharlal Nehru New College of Engineering, Shivamogga, India[2-4]

**Abstract:** This study assesses medical images, particularly Computed Tomography (CT) scans, for the early detection of lung cancer using processing the image, machine learning, and modern technology. The study highlights how raising patient survival rates depends on early-stage detection. Getting accurate standard performance is the primary goal. The methodology involves several processes, including dataset acquisition, data augmentation, pre-processing, selection of features, extraction of features, and CNN implementation. The outcomes of the trial indicate the precision with which our proposed technique works and how it could improve medical imaging in the existing clinical context for prevention and the therapy for lung cancer.

**Keywords:** Lung Cancer (LC), CT scan images, Convolutional Neural Networks (CNN)

## I. INTRODUCTION

One illness that may be devastating is lung cancer. [7]. It is among the worst illnesses of the modern era and has been the primary cause of death globally for the last several decades [1]. Lung cancer destroys more individuals annually than all other types of cancer combined. Both men and women are impacted by the same lethal disease. Following identification of lung cancer, the patient's life expectancy is incredibly short [9]. The sickness gets worse in stages; it begins in the small tissue and spreads to different sections of the lung via a process known as metastasis. It is the unregulated growth of undesirable cells in the lungs. 12,203 people (7130 males and 5073 women) are projected to have had lung disease in 2016, and 8839 of their deaths were due to lung cancer [11].

With millions of new cases and fatalities each year, lung cancer remains the primary the reason behind lung cancer growth rate and death globally. Although smoking is the primary cause in addition to lung cancer, there are non-smoking causes as well. Over 12,000 fatalities from lung cancer would have been avoided If half of individuals in grave danger had received screening for the illness [3]. According to World Health Organization (WHO) reports that cancer is the primary 9.6 million fatalities globally, with around 9.6 million deaths from the disease in 2018. With 1.76 million deaths (18.4%) and 2.09 million cases (11.6%) from the disease, the most typical kind of cancer is lung cancer typical; type of cancer. By 2020, cancer is predicted to be the reason behind of 12 million deaths worldwide, with lung cancer having a pitiful 16% 5-year survival rate. Conversely, however, early nodule discovery and therapy can raise the likelihood that lung cancer will survive [8].

Usually, the symptoms show elevated in the latter stages. Therefore, It is challenging to recognize at this early stage. Its chances of survival increase with correct identification. It is estimated that 85% of occurrences of lung cancer are caused by smoking. Non-smokers are less prone to get this sickness than smokers are [2]. Smoke that enters the lungs damages the lung tissue. Lung cancer in non-smokers can be triggered by radon radiation, air pollution, second hand smoking, lifestyle decisions, and other things. Hereditary factors can potentially result in lung cancer [5].

Breathing becomes more challenging because of tumours forming on lung cells, a frequent sign of lung cancer that has an impact on people everywhere. It is the result of lung cells proliferating quickly, which causes a malignant tumor to grow. Smoking raises the risk factor for this disease because it contains chemicals and carcinogens that cause cancer. The most typical warning signs and manifestations of this deadly sickness are bloody coughing, peculiar coughing, dyspnoea, weariness, appetite loss, weight loss, and, in the later stages, metastasis—the disease spreading to other body areas [4].

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024
DOI: 10.17148/IJARCCE.2024.134153

It is unregulated tissue magnification that leads to lung cancer. The cells that give rise to primary cancer are the starting point for secondary (metastatic) cancer when it moves from another area of the body to the lungs. Both benign and malignant (cancerous) lung cancers are possible. They are ordered based on ranging from numbers and cells they come from. Cancer of either Grade I or II is referred to as lower-grade cancer. In other cases, tumours classified as III and IV are considered higher-grade malignancies. The human body is composed of two different types of cells. Normal cells are often small and contained, but cancer-affected cells are beginning to form and is readily differentiated from them. The cells seem very different from regular cells. These cells divide more quickly and are more likely to spread. They are categorized as either poorly distinguished or high-grade [1].

Lung cancer is identified by the development of tissue clumps inside the lung, or a nodule. Lung nodules can frequently be categorized as juxta-pleural, well-circumscribed, pleural tail, or vascularised depending on their specific location within the lung [8]. Not every lung nodule has to be cancerous. There are two separate staging for the advancement of lung cancer: the number system such as Stage 1, Stage 2, Stage 3, Stage 4 and the TNM (tumor, nodes, metastases) method. The size of tumor determines Stage I and II. Stage IV indicates that the malignant growth has spread to other body areas, including the brain and liver, while Stage 3 only demonstrates lymph node involvement [3].

The most well-liked and extensively used pathological test type for diagnosis is the CT scan. To provide a three-dimensional assessment of the lesion, crisp, high-contrast lung pictures are required [1]. Compared to MRI and X-ray reports, those from CT scans are less noisy [9]. As of right now, CT imaging is taken to be the best imaging modality for investigating and early diagnosis of lung nodules. The continuous cross-sectional images produced by a CT scan of the lung are utilized to confirm the malignant the lung's condition. These images were taken from variety of angles [8].

The skilled physicians determine the phases of the malignancy and make a professional diagnosis. The course of treatment includes targeted therapy, radiotherapy, chemical treatments to kill or stop cancerous cells from proliferating, and a few surgical procedures. These analyses involve painful bodily parts and demand a significant amount of and money. Therefore, utilizing a variety of image processing methods to expedite this process [9].

Using gray scale CT scan pictures, this investigation has proven an automated approach for cancer diagnosis. The input consisted of photographs of lung cancer, and the output images were produced by applying pre- and post-processing methods for processing medical images once the region had been narrowed. Pre-processing consists of segmentation, filter operation, and enhancement. The post-processing consists of feature selection and identification [1].

## II. LUNG CANCER PREDICTION APPROACHES

The many projects completed in the designated area are listed below. Every paper is discussed along with its methodologies, benefits, shortcomings, and concise methodology.

### 2.1 "A Comparative Study of Lung Cancer Detection Using Supervised Neural Network"

This study compares the prognosis of lung cancer using supervised neural networks. The authors pre-process CT scan pictures and apply machine learning and biomedical image processing techniques, including segmentation, Random Forest, and SVM algorithm application. The greatest results are attained using SVM classification; the suggested strategy has a 94.5% accuracy rate in lung cancer detection. The study emphasizes how important early detection is to effective treatment, especially in places where lung cancer risk has grown due to pollution and smoking.

### 2.2 "A Study on Early Prediction of Lung Cancer Using Machine Learning Techniques"

Using machine learning techniques for detecting lung cancer (LC) prediction is examined in this article. Special consideration is given to the feature selection, pre-processing, workflow strategy, data collecting, and performance metrics of various machine learning techniques. The goal of the study is to improve the effectiveness and precision of LC prediction while providing useful information for future developments in machine learning-based cancer detection.

### 2.3 "Automatic Detection of Lung Cancer Identification using ENNPSO Classification"

The three primary elements of the proposed automated lung cancer detection approach are the Elman Neural Network (ENN) with Particle Swarm Optimization (PSO) for classification, an Active Contour approach (ACM) for segmentation, and an Enhanced Kaun Filter (IKF) for pre-processing CT scan images. The method outperforms traditional classifiers like SVM, RBFN, and ANFIS, with an accuracy rate of 93%. The outcomes indicate the efficacy of the ENNPSO approach in differentiating lung cancer from DICOM images, potentially leading to early detection and diagnosis.

  

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024
DOI: 10.17148/IJARCCE.2024.134153

### 2.4 "Early Stage Lung Cancer Prediction Using Various Machine Learning Techniques"

The proposed project will carefully examine many methods of machine learning techniques for patient early-stage lung cancer prediction.The pipeline includes data pre-processing steps like feature selection, categorical data encoding, and handling missing data. The experiment evaluates K-Nearest Neighbor, Random Forest, Support Vector Machine, Neural Networks, and a Voting Classifier; the latter attains the highest accuracy of 99.5%. In the study's conclusion, the Voting Classifier is suggested as the most accurate lung cancer prediction model in its early stages. Possible directions for further development, like examining other models and pre-processing techniques, are also mentioned.

### 2.5 "Early Stage Lung Cancer Diagnosis using ANN Classifier"

The recommended strategy makes application of processing image methods along in addition to an ANN classifier, artificial neural networks (ANN) to identify lung cancer in its early stages. The steps in the approach include pre-processing CT scans, segmenting lung areas, eliminating relevant nodules, extracting features, and finally applying an ANN for classification. The method demonstrated promising results and provided a valuable tool for successful lung cancer identification, with an overall accuracy of 95.6%. Future research may broaden the applicability to various organs and imaging modalities with a focus on improving precision and integrating a larger range of data.

### 2.6 "Lung Cancer Classification and Prediction Using Machine Learning and Image Processing"

The work provides a sound approach that blends machine learning and image processing methods for the classification and prediction of lung cancer. It outlines the procedures, which include obtaining images, segmenting the images using K-means, pre-processing with a geometric mean filter, and finally categorizing the outcomes using machine learning methods. The study shows how important is to diagnose lung cancer precisely and how Artificial Neural Networks (ANN) have higher prediction accuracy. When everything is said and done, the study promotes the application of state-of-the-art machine learning methods to enhance lung cancer detection.

### 2.7 "Segmenting Nodules of Lung Tomography Image with Level Set Algorithm and Neural Network"

The work described in the study aims develops a method for segmenting images that is applicable to identify lung nodules on computed tomography (CT) images. The suggested approach consists of a median filter-based pre-processing stage, segmentation using the level set method (LSM), feature extraction (diameter, centreline, and area), and classification using artificial neural networks (ANNs). Utilizing a Lung Image Database Consortium dataset (LIDC), the study's performance analysis demonstrates that LSM performs better for bigger nodules. The suggested technique improves the field of utilizing computer-aided diagnosis in the context of lung cancer early detection.

### 2.8 "Image processing and machine learning for lung cancer detection in healthcare"

After the supplied CT scan picture has been pre-processed, several machine learning and image processing algorithms, such as grayscale conversion, noise reduction, and binarization, are employed in this article to characterise lung cancer and its phases. For pre-processing stages, the segmentation and median filter yield precise results. For pre-processing stages, the segmentation and median filter yield precise results. For grouping purposes, the system's Support Vector Machine (SVM) classifier categorises positive and negative data related to lung cancer images.

### 2.9 "CT Scan Images for the Detection of Lung Cancer".

This study shows how segmentation and SVM may be used to categorize nodules as benign or malignant in order to identify cancerous nodules from lung CT scan pictures. From the identified cancer nodes, characteristics including area, perimeter, centroid, diameter, eccentricity, and mean intensity of pixels were extracted. A support vector machine was trained using the extracted features, and a trained model was produced. It has a 92% accuracy rate in identifying cancer, which is greater than the 86.6% accuracy rate of the present model and classifier.

### 2.10 "Lung Cancer Detection Using the KNN Algorithm and SVM Classifier"

In this work, medical data—which may include imaging or clinical details—is analysed using the Support Vector Machine (SVM) and k-Nearest Neighbor (KNN) algorithms to precisely predict lung cancer. In this work, image processing techniques are used to identify early-stage lung cancer in CT scan pictures. The ROI of the lungs is divided once the CT filter image has been prepared. Separate waveform for picture pressure, the transform is linked, and a GLCM is used to extract the highlights. The findings are used by an SVM classifier to determine whether or not the lung picture is carcinogenic. The LIDC dataset is used to assess the SVM classifier.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134153

## III. PROPOSED METHODOLOGY

We conducted five phases to establish if the cancer was likely to arise. As shown in Figure 1, first gather the dataset, then pre-process the images, train the CNN algorithm, create the model file, and then test the images.

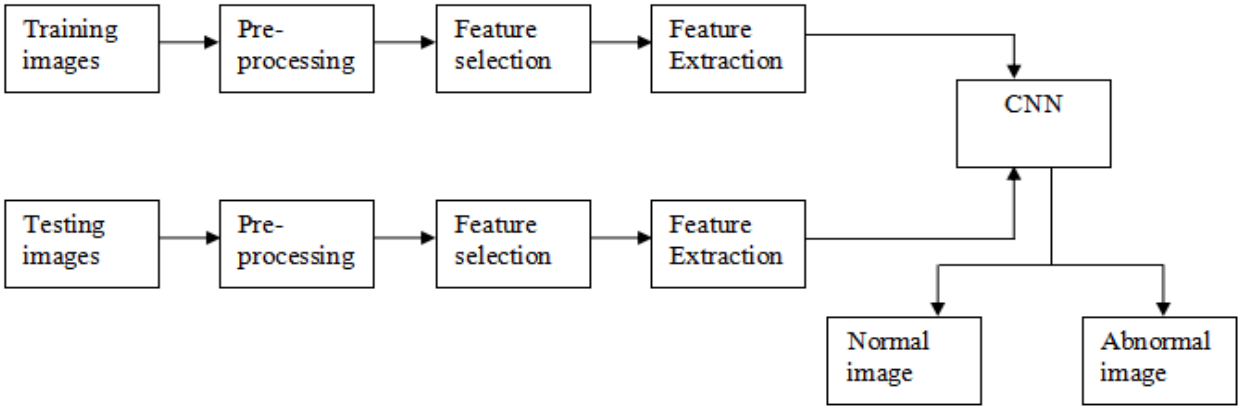


Figure (1): The block diagram of the proposed work

Acquiring lung CT images from different medical facilities, such as clinics and hospitals, is required to create an extensive dataset. It is important to make sure that the images in this collection depict both malignant and non-cancerous lung illnesses. After they are acquired, the photos undergo meticulous preprocessing. To achieve standardization and encourage uniformity throughout the dataset, resizing to a common dimension—typically 224 × 224 pixels—is required. Additionally, the process is made more interpretable by using normalization procedures to ensure consistency in color and brightness levels. Because the dataset is robustly and diversely provided via augmentation methods like flipping and rotation, the model is able to generalize successfully. Like a keen detective, the Convolutional Neural Network (CNN) algorithm is then taught. Supplemented with labels specific to each lung picture. When presented with labels indicating whether or not each individual lung picture.

## IV. CNN ARCHITECTURE

Convolutional Neural Networks (CNNs) have great potential in deep learning, particularly in tasks related to image processing and analysis. CNNs are highly proficient in recognising intricate patterns found in images because they are composed of several layers, including pooling, fully connected, and convolutional layers. The brains of a CNN are its convolutional layers, which apply filters to the input image and successfully extract significant features like edges, textures, and shapes. As these convolutional layers develop, the recovered properties are improved and refined.

Convolutional layers are followed by the pooling methods on the feature maps. Pooling layers contribute to a reduction in the spatial dimensions of feature maps by retaining the most significant information while down sampling the feature maps. This down sampling process increases computing efficiency by distilling the data and focusing on the most crucial elements. After the feature maps are processed by layer pooling, the output is processed by one or more fully linked layers. These layers enable the CNN to handle image input for classification or prediction needs. The feature maps' data is combined by the completely linked layers to facilitate higher-level reasoning and decision-making. In the end, this aids CNN in precisely predicting or classifying based on the input image. CNNs are necessary for any activity that involves picture identification and processing, as they employ the layers' hierarchical structure to efficiently extract, alter, and interpret visual information.

The supplied images are downsized to (224, 224, 3) to match the RGB colour channels, width, and height. The Conv2D layer starts the process by applying 128 filters with ReLU activation and a size of (3, 3). A (2, 2) window size is used by the succeeding MaxPooling2D layers to minimize spatial dimensions. Subsequently, a second Conv2D layer uses 64 (3, 3)-sized filters with ReLU activation to maximize feature extraction. Another MaxPooling2D layer comes after this. A second layer of MaxPooling2D captures even more complexity, and it is followed by a second layer of Conv2D with 32 filters and ReLU activation. In order to prepare the data for the fully connected layers, the multi-dimensional output is subsequently flattened into a one-dimensional array. Next, there was a dense, thick coating.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134153

## V. COMPARATIVE ANALYSIS

Table (1): Comparative study table

| Sl.no | Algorithm | Accuracy | Loss |
|---|---|---|---|
| 1 | CNN | 100% | 38% |
| 2 | AlexNet | 45% | 96% |
| 3 | Inception V3 | 70% | 12.7% |

## VI. RESULT AND ANALYSIS

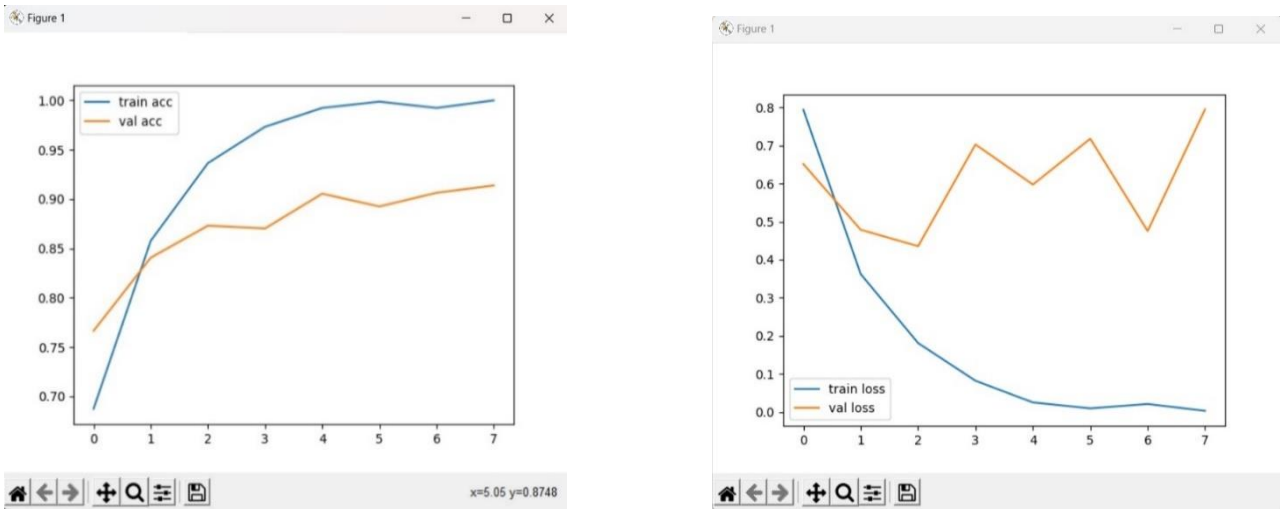


Figure (2): CNN graph loss and accuracy

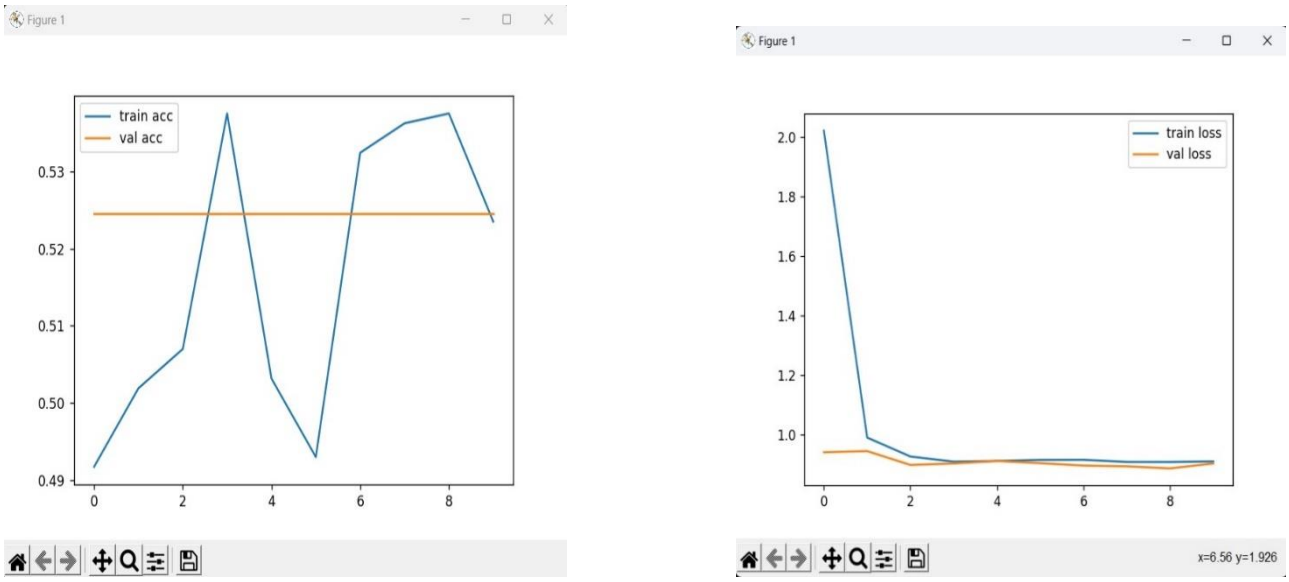


Figure (3): AlexNet graph loss and accuracy

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134153

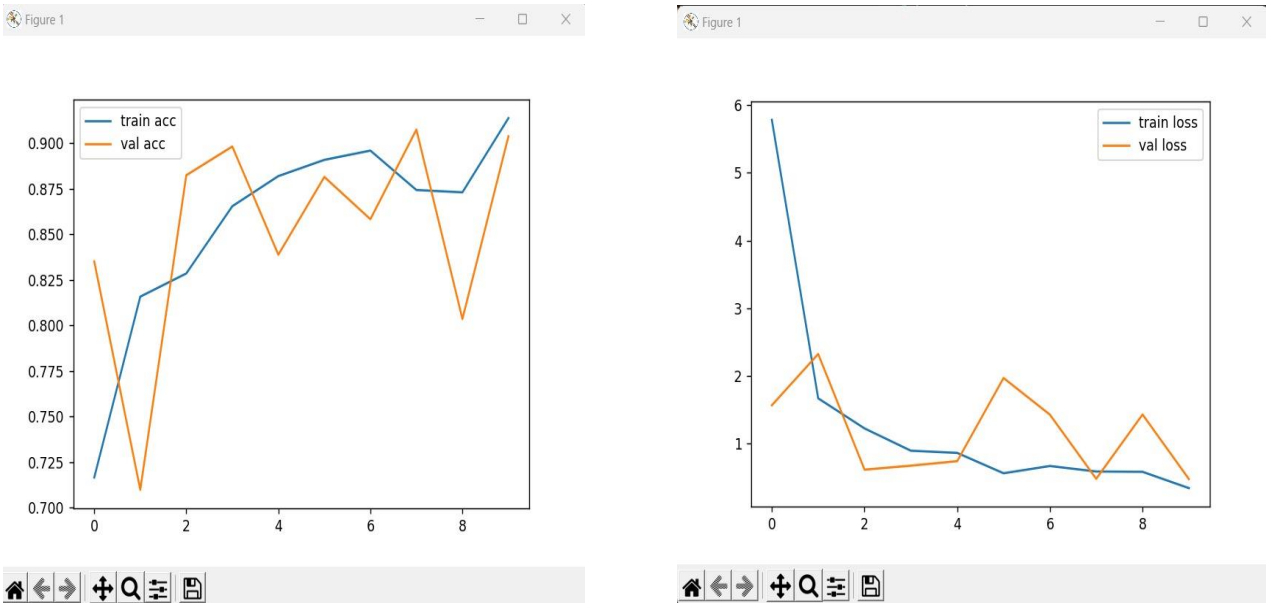


Figure (4): Inception V3 loss and accuracy

After normalisation, the input pictures are set to (224, 224, 3) for the RGB colour channels, height, and width. Applying 128 filters with ReLU activation in size (3, 3) is how the Conv2D layer begins the process. Next-generation MaxPooling2D layers reduce spatial dimensions using a (2, 2) window size. The next step is a second Conv2D layer that further enhances feature extraction by employing 64 (3, 3)-sized filters and ReLU activation. This is followed by another layer of MaxPooling2D. Further intricacy is captured by a second MaxPooling2D layer that is followed by an extra Conv2D layer with 32 filters activated by ReLU. The data is then prepared for the fully linked layers by flattening the multi-dimensional output into a one-dimensional array. Following it, there was a dense layer with 512 neurons and ReLU activation.

## VII. CONCLUSION

Medical imaging analysis is essential to the detection and treatment of disorders such as lung cancer. This paper presents a unique approach that combines feature selection with a Convolutional Neural Network (CNN) classifier to enhance the identification of lung cancer using CT images. Finding lung cancer as soon as feasible and with great precision is the major objective of this investigation. Manual lung cancer diagnosis is time-consuming, subjective, and often dependent on the expertise of laboratory personnel. Although automated methods get results faster, they typically involve complex processes.

This study aims to improve diagnostic precision and reduce processing time by simplifying and streamlining current approaches. Though advancements have been made, more may be done, particularly in terms of sharpening the accuracy of grayscale photo analysis. Future research endeavors might focus on addressing problems related to variations in interpretation among physicians, both within and between specialties. By addressing these problems, the recommended approach may enhance the detection of lung cancer and contribute to improved patient care.

## REFERENCES

[1] Fahad Ahmed, Md. Zahidul Islam, Ariful Hoque, and A.K.M. Ashek Farabi, "Automated Detection of Lung Cancer Using CT Scan Images," IEEE Access, pp. 5-7, 2020.

[2] Kyamelia Roy, Sheli Sinha Chaudhury, Madhurima Burman, Ahana Ganguly, Chandrima Dutta, Sayani Banik, and Rayna Banik, "A Comparative Study of Lung Cancer Detection Using a Supervised Neural Network," IEEE Access, pp. 978-1-7281-0070-8, 2019.

[3] Ms. V. Nisha Jenipher, Dr. S. Radhika, "A Study on Early Prediction of Lung Cancer Using Machine Learning Techniques," IEEE Access, Third International Conference on Intelligent Sustainable Systems [ICISS 2020], pp. 978-1-7281-7089-3, 2020.

[4] B. Hemalatha, S. Yuvaraj, K.V. Kiruthikaa, and V. Viswanathan, "Automatic Detection of Lung Cancer Identification Using ENNPSO Classification," IEEE Access, pp. 2020.

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134153

[5] Chinmayi Thallam, Aarsha Peruboyina, Sagi Sai Tejasvi Raju, and Nalini Sampath, "Early Stage Lung Cancer Prediction Using Various Machine Learning Techniques," IEEE Access, Fourth International Conference on Electronics, Communication, and Aerospace Technology (ICECA-2020), pp. 978-1-7281-6387-1, 2020.

[6] Mr. Shailesh S. Bhise, Prof. S. R. Khot, "Early Stage Lung Cancer Diagnosis using ANN Classifier" ,IEEE access, International Conference on Artificial Intelligence and Smart Systems (ICAIS-2021),pp.978-1-7281-9537-7,2021.

[7] Sharmila Nageswaran, G. Arunkumar, Anil Kumar Bisht, Shivlal Mewada, J. N. V. R. Swarup Kumar , Malik Jawarneh, and Evans Asenso, "Lung Cancer Classification and Prediction Using Machine Learning and Image Processing," Hindawi , vol. 8, pp. 2022.

[8] Soon Yee Chong , Min Keng Tan, Kiam Beng Yeo , Mohd Yusof Ibrahim , Xiaoxi Hao , Kenneth Tze Kin Teo , "Segmenting Nodules of Lung Tomography Image with Level Set Algorithm and Neural Network", IEEE access,7th Conference on Systems, Process and Control (ICSPC 2019),pp from 13-14,2019.

[9] Wasudeo Rahane, Himali Dalvi, Yamini Magar, Anjali Kalane, and Satyajeet Jondhale, "Lung Cancer Detection Using Image Processing and Machine Learning HealthCare," IEEE Access, International Conference on Current Trends towards Converging Technologies, pp. 978-1-5386-3702-9, 2018.

[10] Suren Makajua , P.W.C. Prasad*a, Abeer Alsadoona , A. K. Singhb, and A. Elchouemi, "Lung Cancer Detection using CT Scan Images," Elsevier, 6th International Conference on Smart Computing and Communications, pp. 7-8, 2018.

[11] R. Satishkumar, K. Kalaiarasan, A. Prabhakaran, and M. Aravind, "Detection of Lung Cancer using SVM Classifier and KNN Algorithm," IEEE Access, International Conference on Systems Computation Automation and Networking, pp. 978-1-7281-1524-5, 2019.