A Project Report on

Lung Cancer Detection using Machine Learning

A Dissertation submitted to JNTU Hyderabad in partial fulfillment of the academic requirements for the award of the degree.

Bachelor of Technology

in

Computer Science and Engineering

Submitted by

P. THARUN SAI (20H51A0522)
C. GANESH (20H51A0560)
K. SHIVA ABHIGNA (20H51A05H3)

Under the esteemed guidance of

Mr. A. Vivekanand

(Associate Professor)

Department of Computer Science and Engineering

CMR COLLEGE OF ENGINEERING & TECHNOLOGY

(UGC Autonomous)

*Approved by AICTE *Affiliated to JNTUH *NAAC Accredited with A^+ Grade

KANDLAKOYA, MEDCHAL ROAD, HYDERABAD - 501401.

2020- 2024

CMR COLLEGE OF ENGINEERING & TECHNOLOGY

KANDLAKOYA, MEDCHAL ROAD, HYDERABAD – 501401

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the Major Project report entitled **" LUNG CANCER DETECTION USING MACHINE LEARNING"** being submitted by P. Tharun Sai (20H51A0522), C. Ganesh (20H51A0560) and K. Shiva Abhigna (20H51A05H3) in partial fulfillment for the award of **Bachelor of Technology in Computer Science and Engineering** is a record of bonafide work carried out his/her under my guidance and supervision.

The results embodied in this project report have not been submitted to any other University or Institute for the award of any Degree.

Mr. A. Vivekanand
Associate Professor
Dept. of CSE

Dr. Siva Skandha Sanagala
Associate Professor & HOD
Dept. of CSE

EXTERNAL EXAMINER

ACKNOWLEDGEMENT

With great pleasure we want to take this opportunity to express my heartfelt gratitude to all the people who helped in making this project work a grand success.

We are grateful to **Mr. A. Vivekanand, Associate Professor**, Department of Computer Science and Engineering for his valuable technical suggestions and guidance during the execution of this project work.

We would like to thank **Dr. Siva Skandha Sanagala,** Head of the Department of Computer Science and Engineering, CMR College of Engineering and Technology, who is the major driving forces to complete my project work successfully.

We are very grateful to **Dr. Ghanta Devadasu**, Dean-Academics, CMR College of Engineering and Technology, for his constant support and motivation in carrying out the project work successfully.

We are highly indebted to **Major Dr. V A Narayana,** Principal, CMR College of Engineering and Technology, for giving permission to carry out this project in a successful and fruitful way.

We would like to thank the **Teaching & Non- teaching** staff of Department of Computer Science and Engineering for their co-operation

We express our sincere thanks to **Shri. Ch. Gopal Reddy**, Secretary, CMR Group of Institutions, and **Shri. Ch. Abhinav Reddy**, CEO, CMR Group of Institutions for their continuous care and support.

Finally, we extend thanks to our parents who stood behind us at different stages of this Project. We sincerely acknowledge and thank all those who gave support directly and indirectly in completion of this project work.

P. THARUN SAI 20H51A0522
C. GANESH 20H51A0560
K. SHIVA ABHIGNA 20H51A05H3

TABLE OF CONTENTS

List of Figures

List of Tables

ABSTRACT

Lung cancer is one of the leading causes of death worldwide and ranks among the primary cause of death on global scale. Early detection of this disease increases the chances of survival. Computer-Aided Detection (CAD) has been used to process or to create CT images and even X-rays of the lungs to determine whether an image has traces of cancer or presence of cancer nodules in the images. This project presents an image classification by the combination of a Neural Network (CNN) algorithm and Support Vector Machine (SVM). This algorithm is capable of automatically classifying and analyzing each lung image to check if there is any presence of cancer cells or not. CNN is easier to train and has fewer parameters compared to a fully connected network. we came out with CNN-SVM because it gives good performance compared with other results. This method helps for better merit and its ability to classify lung cancer in CT images or X-rays accurately and detect cancer nodule effectively.

CHAPTER 1
INTRODUCTION

1. INTRODUCTION

1.1. Introduction

Lung cancer detection refers to the process of identifying the presence of lung cancer or the likelihood of its occurrence in an individual's lungs. This detection can be achieved through various medical techniques and diagnostic tools and even with machine and deep learning techniques. Lung cancers usually are grouped into two main types called small cell and non-small cell (non-small cell includes adenocarcinoma and squamous cell carcinoma). These types of lung cancer grow differently and are treated differently. Non-small cell lung cancer is more common than small cell lung cancer [1].

Lung cancer is one of the causes of cancer deaths. It is difficult to detect because it arises and shows symptoms in final stage. However, mortality rate and probability can be reduced by early detection and treatment of the disease. Best imaging technique CT imaging are reliable for lung cancer diagnosis because it can disclose every suspected and unsuspected lung cancer nodule [2], However, variance of intensity in CT scan images and anatomical structure misjudgment by doctors and radiologists might cause difficulty in marking the cancerous cell [3]. Recently, to assist radiologists and doctors detect the cancer accurately computer Aided Diagnosis has become supplement and promising tool [4]. There have been many systems developed and research going on detection of lung cancer. However, some systems do not have satisfactory accuracy of detection and some systems still has to be improved to achieve highest accuracy tending to 100%. Image processing techniques and machine learning techniques has been implemented to detect and classify the lung cancer.

1.2. Problem Statement

Machine and deep learning methods are used widely in the medical and healthcare field for monitoring, detecting, classifying and predicting diseases. Our system deals with CNN-SVM architecture which removes and eliminates useless information that negatively impacts accuracy. This is accomplished in the CNN architecture's pooling step. the last layer of the CNN is replaced with the SVM as a binary classifier and detect cancer or non-cancer nodules which are two different classes or categories.

1.3 Research Objective

In this project we are mainly addressing 3 objectives that are:

1.2.1. The primary objective of using machine learning (ML) for lung cancer detection is to enhance the accuracy and efficiency of identifying potential cases of lung cancer from medical data.

1.2.2. The objective of this project is to develop a hybrid Convolutional Neural Network (CNN) algorithm and Support Vector Machine (SVM) model.

1.2.3. CNN is known to have the best performance when using large amounts of data lacking in medical imaging due to several factors such as ethics and lackof well-labelled data.

1.4 Scope of the Project

The analysis and study of lung cancer/diseases has been the most intriguing investigation zone of medical experts from early days to the present day. To address this concern, a diagnosis system like this can only help diminish the odds of getting risk to human lives. The scope for this proposed system mainly works in medical field and it will be very useful to the medical experts.

CHAPTER 2
BACKGROUND WORK

2. BACKGROUND WORK

2.1 A summary of reference papers

In this section we have studied various implementations of lung cancer detections & we summarized our findings that we concluded by researching & referencing various papers. They are as below:

Mokhled S. Al-Tarawneh (August,2012) [5]. Lung cancer is a disease of abnormal cells multiplying and growing into a tumor. Cancer cells can be carried away from the lungs in blood, or lymph fluid that surrounds lung tissue. Lymph flows through lymphatic vessels, which drain into lymph nodes located in the lungs and in the center of the chest. Lung cancer often spreads toward the center of the chest because the natural flow of lymph out of the lungs is toward the center of the chest. Metastasis occurs when a cancer cell leaves the site where it began and moves into a lymph node or to another part of the body through the blood stream [6]. Several researchers have proposed and implemented detection of lung cancer using different approaches of image processing and machine learning.

Aggarwal, Furquan and Kalra [7] proposed a model that provides classification between nodules and normal lung anatomy structure. The method extracts geometrical, statistical and gray level characteristics. LDA is used as classifier and optimal thresholding for segmentation. The system has 84% accuracy, 97.14% sensitivity and 53.33% specificity. Although the system detects the cancer nodule, its accuracy is still unacceptable. No any machine learning techniques has been used to classify and simple segmentation techniques is used. Therefore, combination of any of its steps in our new model does not provide probability of improvement.

Jin, Zhang and Jin [8] used convolution neural network as classifier in his CAD system to detect the lung cancer. The system has 84.6% of accuracy, 82.5% of sensitivity and 86.7% of specificity. The advantage of this model is that it uses circular filter in region of interest (ROI) extraction phase which reduces the cost of training and recognition steps. Although, implementation cost is reduced, it has still unsatisfactory accuracy. Sangamithraa and Govindaraju [9] uses K mean unsupervised learning algorithm for clustering or segmentation. It groups the pixel dataset according to certain characteristics. For classification this model implements back propagation network. Features like entropy, correlation, homogeneity, PSNR, SSIM are extracted using gray-level co-occurrence matrix (GLCM) method. The system has accuracy of about 90.7%. Image preprocessing median filter is used for noise removal which can be useful for our new model to remove the noise and improve the accuracy.

Roy, Sirohi, and Patle [10] developed a system to detect lung cancer nodule using fuzzy interference system and active contour model. This system uses gray transformation for image contrast enhancement. Image binarization is performed before segmentation and resulted image is segmented using active contour model. Cancer classification is performed using fuzzy inference method. Features like area, mean, entropy, correlation, major axis length, minor axis length are extracted to train the classifier. Overall, accuracy of the system is 94.12%. Counting its limitation it does not classify the cancer as benign or malignant which is future scope of this proposed model.

Ignatious and Joseph [11] developed a system using watershed segmentation. In preprocessing it uses Gabor filter to enhance the image quality. It compares the accuracy with neural fuzzy model and region growing method. Accuracy of the proposed is 90.1% which is comparatively higher than the model with segmentation using neural fuzzy model and region growing method. The advantage of this model is that it uses marker-controlled watershed segmentation which solves over segmentation problem. As a limitation it does not classify the cancer as benign or malignant and accuracy is high but still not satisfactory. Some changes and contribution in this model, have probability of increasing the accuracy to satisfactory level.

2.2 lung disease detection based on IoT in conjunction with Machine Learning:

2.2.1 Introduction

Lung cancer detection using the Internet of Things (IoT) involves the use of connected sensors and devices to continuously monitor lung health parameters, such as oxygen levels and lung function [22].

1. **Respiratory Monitoring:** IoT devices can monitor parameters such as respiratory rate, oxygen saturation, and lung function. Deviations from normal values may indicate respiratory issues.
2. **Environmental Sensors:** IoT sensors can measure air quality and pollutants, which can be linked to respiratory problems. For instance, elevated levels of air pollutants may contribute to lung diseases.
3. **Wearable Devices:** Wearable IoT devices can track physical activity, heart rate, and even lung sounds. Changes in these metrics can signal potential lung health concerns.
4. **Spirometers:** IoT-connected spirometers can measure lung capacity and airflow rates, assisting in the early detection of lung diseases like chronic obstructive pulmonary disease (COPD).
5. **Pulse Oximeters:** These devices measure oxygen saturation in the blood, which is crucial for detecting lung diseases like pneumonia and other respiratory conditions.
6. **Smart Inhalers:** IoT-enabled inhalers can help patients manage chronic respiratory conditions by tracking medication usage and adherence [22].

2.7.2 Merits and Demerits:

Merits:

- IoT enables continuous monitoring of patients, providing real-time data on lung health and early detection of abnormalities.
- By combining IoT and machine learning, early detection and personalized treatment for lung cancer can be enhanced, leading to better patient care.

Demerits:

- IoT devices require regular calibration and maintenance to ensure data accuracy. failure to do so can result in incorrect diagnoses.

2.2.3 Implementation:

- **IoT Devices for Data Collection:** IoT devices gather patient data like vital signs and environmental factors.
- **Machine Learning Algorithms for Analysis:** Machine learning algorithms analyze this data to find patterns indicating lung cancer.
- **Alerts:** Real-time monitoring triggers alerts for healthcare providers or patients if signs of lung cancer are detected.
- **Personalized Treatment:** Tailored treatment plans based on individual characteristics and data analysis are created for better outcomes.
- **Continuous Improvement:** The system learns from data over time to improve accuracy and effectiveness.

2.3 Random Forest algorithm for lung cancer classification:

2.3.1 Introduction

Early detection is crucial for successful lung cancer treatment. Machine learning algorithms like Random Forests offer promising possibilities. By analyzing patient data like demographics, smoking history, and imaging features, Random Forests can estimate an individual's risk of lung cancer. This allows doctors to personalize screening strategies, directing high-risk patients towards more frequent checks while potentially sparing low-risk individuals from unnecessary procedures. This data-driven approach holds promise for improving lung cancer detection and patient outcomes [23].

2.3.2 Merits and Demerits:

Merits:

- Random Forests are less prone to overfitting compared to other machine learning algorithms, such as decision trees. This is because they aggregate multiple decision trees, each trained on a random subset of the data and features, which helps to generalize well to unseen data.
- Random Forests can efficiently handle large datasets with high-dimensional feature spaces, making them suitable for analyzing medical imaging data and other complex datasets often encountered in lung cancer detection.
- They are capable of capturing complex relationships between input features and target classes, leading to accurate predictions.

Demerits:

- Random Forests are considered black box models, Understanding the decision-making process behind Random Forests can be challenging, particularly for complex datasets.
- Random Forests have several hyperparameters that need to be tuned for optimal performance, such as the number of trees, tree depth, and minimum leaf size. Finding the right combination of hyperparameters can require extensive experimentation and computational resources.

2.3.3 Implementation:

- **Data Collection:** Gather a dataset consisting of lung images or other relevant features. This dataset should include both positive cases (lung cancer patients) and negative cases (healthy individuals or non-cancerous lung conditions).
- **Data Preprocessing:** Preprocess the dataset to ensure data quality and consistency. This may involve steps such as resizing images, normalizing pixel values, and handling missing or noisy data.
- **Data Splitting:** Split the dataset into training and testing sets. The training set will be used to train the Random Forest classifier, while the testing set will be used to evaluate its performance.
- **Model Training:** Train the Random Forest classifier using the training data. The classifier will learn to distinguish between lung cancer and non-cancer cases based on the extracted features.
- **Hyperparameter Tuning:** Optimize the hyperparameters of the Random Forest classifier to improve its performance. Hyperparameters include the number of trees in the forest, maximum tree depth, minimum samples per leaf, and other parameters that control the behavior of the algorithms
- **Model Evaluation:** Evaluate the trained Random Forest classifier using the testing data. Common

evaluation metrics for binary classification tasks include accuracy, precision, recall, F1 score, and area under the ROC curve.

Table 2.1: Comparison table based on various research papers studied

Reference	Author	Title	Year of Publishing	Results
[12]	Saba	Automated lung nodule detection and classification based on multiple classifiers voting.	2019	Accuracy-96.4%
[13]	Firmino et al.	Computer-aided detection (CADe) and diagnosis (CADx) system for lung cancer with likelihood of malignancy.	2016	Accuracy-92%
[14]	S.M. Naqi, M. Sharif, I.U. Lali	A 3D nodule candidate detection method supported by hybrid features to reduce false positives in lung nodule detection.	2019	Accuracy-86.9%
[15]	Asuntha and Srinivasan	Deep learning for lung Cancer detection and classification.	2019	Accuracy-75.62%
[16]	S.A. Khan, M. Nazir, M.A. Khan, T. Saba, K. Javed, A. Rehman	Lung nodule detection framework from computed tomography images using support vector machine.	2017	Accuracy-86%
[17]	D. Kumar, A. Wong, D.A. Clausi	Lung nodule classification using deep features in CT images.	2015	Accuracy-89%

[18]	Muzammil etal.	Ensemble Learning Based Fusion, Pulmonary nodule classification using feature and ensemble learning-based techniques	2021	Accuracy-96%
[19]	Bansal et al.	Deep3DSCcan: Deep esidual network and orphological descriptor-based framework for lung ancer classification and 3D segmentation, IET Image Process.	2020	Accuracy-92.7%
[20]	Shah et al.	A lung nodule classification using deep learning, Transfer Learning VGG16 and 19	2020	Accuracy-94%
[21]	Guo et al	Novel computer-aided lung cancer detection based on convolutional neural network-based and feature-based classifiers using metaheuristics.	2021	Accuracy-89%

CHAPTER 3
PROPOSED SYSTEM

3. PROPOSED SYSTEM

3.1. Objective of Proposed Model

Advance diagnosis of lung cancer can reduce mortality. It is main principal to catch cancer early, prevent its development, and eliminate it early before it starts to grow rapidly. Technology and deep learning are broadly used in medicine to monitor, detect, classify, and predict diseases. Our system includes a CNN-SVM architecture that removes and deletes redundant information that affects accuracy. This is done in the step of integrating the CNN architecture and in this CNN architecture the last layer is replaced by SVM classifier in which the flattened output from CNN architecture is sent into that last layer that is SVM, In this layer the classification takes place that is to determine whether the test sample is normal or abnormal that is cancerous. The integration of CNNs and SVMs offers several advantages. Firstly, it allows for a more comprehensive analysis of lung scans, capturing both local and global features essential for accurate cancer detection. Secondly, the hybrid system combines the superior feature extraction capabilities of CNNs with the robust classification abilities of SVMs, resulting in a more effective and efficient lung cancer detection system. this proposed hybrid CNN-SVM system presents a promising approach to lung cancer detection, potentially improving diagnostic accuracy and aiding in the early detection and treatment of this deadly disease. Further research and validation of this system could significantly enhance the capabilities of medical professionals in combating lung cancer.

3.2 Algorithms Used in Proposed Model

In this proposed system for lung cancer detection using machine learning, two key algorithms play crucial role: Neural network CNN for feature extraction purpose, Support Vector Machine (SVM) for classification. Here is how each algorithm contributes to the system:

1. **Convolutional Neural Networks:**

In the proposed system, the CNN plays a crucial role in the initial stages of processing the medical images, specifically the lung scans. Here is how CNN works in this framework:

Feature Extraction: Convolutional Neural Networks (CNNs) are adept at automatically learning and extracting complicated features from images. In the context of lung cancer detection, the CNN is trained on a dataset of lung scans, learning to identify patterns, textures, and structures that may be indicative of cancerous growths or abnormalities within the lungs. These learned features are hierarchical and represent different levels of

abstraction, allowing the CNN to capture both local and global characteristics of the lung images.

Convolutional Layers: The CNN consists of multiple layers, including convolutional layers, pooling layers, and fully connected layers. In the convolutional layers, filters or kernels convolve across the input lung images, extracting features such as edges, textures, and shapes. Through the application of non-linear activation functions like ReLU (Rectified Linear Unit), the CNN enhances its ability to capture complex patterns and representations within the images.

Pooling Layers: Pooling layers follow the convolutional layers and serve to down sample the feature maps generated by the convolutional layers. This reduces the dimensionality of the feature maps while retaining the most important information. Pooling helps in making the CNN more robust to variations in the input images and aids in computational efficiency.

Fully Connected Layers: The output of the convolutional and pooling layers is flattened and fed into fully connected layers. These layers act as classifiers, learning to map the extracted features to the presence or absence of lung cancer. Through a process of training on labeled data (lung scans with known cancer status), the CNN learns to distinguish between cancerous and non-cancerous lung images.

Training and Optimization: During the training phase, the parameters of the CNN, including the weights and biases of the individual layers, are adjusted iteratively using optimization algorithms such as stochastic gradient descent (SGD) or Adam. The objective is to minimize a predefined loss function, which measures the disparity between the predicted outputs of the CNN and the ground truth labels of the training data.

Feature Representation for SVM: Once trained, the CNN serves as a feature extractor, transforming the input lung scans into high-dimensional feature vectors that encode the learned representations. These feature vectors, which capture the salient characteristics of the lung images, are then used as input to the Support Vector Machine (SVM) classifier.

2. Support Vector Machine (SVM):

SVM is one of the prominent machine learning algorithms used for learning techniques used in classification and also for regression problems. Even though, in machine learning concept it is widely used only in classification problems which classify two different classes. the SVM acts as the final decision-maker, using the information provided by the CNN to draw a clear line between healthy and potentially cancerous lung scans. Its goal is to make confident and accurate predictions about the health status of the lungs based on the extracted features. The SVM tries to maximize the distance between this boundary line and the closest points from each group of lung scans. This distance is called the margin. By maximizing this margin, the SVM aims to make the most confident decisions about whether a lung scan is healthy or potentially cancerous.

3.3 Designing

The architecture of lung cancer detection contains various steps that helps to easily understand the flow process of our system they are:

- **Upload Dataset and Data Collection**
- **Splitting the Datasets**
- **Preprocess Data or Data Preprocessing**
- **CNN Model Design and Training**
- **SVM Model Design and Training**
- **Display Accuracies**
- **Model Evaluation and Validation**
- **Comparison graph**

1. Upload Dataset and Data Collection:

Gather a large dataset of lung scans, with each image labeled as either healthy or cancerous. This dataset will be used for training and testing the system and Allow users to upload the dataset containing lung X-ray images.

2. Splitting the datasets:

Split the datasets for training and testing.

3. Data processing:

Preprocess the collected data to ensure uniformity and quality. This may involve tasks such as resizing images to a standard size, normalizing pixel values, and removing noise or artifacts from the images. Data augmentation techniques such as rotation, flipping, and cropping can also be applied to increase the diversity of the training dataset.

4. CNN Model Design and Training:

Design the architecture of the Convolutional Neural Network (CNN) specifically tailored for lung cancer detection. This involves deciding on the number of convolutional layers, pooling layers, and fully connected layers, as well as choosing appropriate activation functions and regularization techniques.

Train the CNN using the preprocessed dataset. During training, adjust the parameters of the network iteratively using optimization algorithms such as stochastic gradient descent (SGD) or Adam. Monitor the training process to ensure that the CNN learns to extract meaningful features from the lung images.

5. SVM Model Design and Training:

Design the Support Vector Machine (SVM) classifier to utilize the feature vectors extracted from the CNN. Choose appropriate kernel functions and regularization parameters for the SVM.

Train the SVM classifier using the feature vectors extracted from the training set. Adjust the hyperparameters of the SVM to optimize its performance in classifying lung images as healthy or cancerous.

6. Display Accuracies:

The Accuracies of any ideal algorithm like SVM, K-Means or any supervised learning algorithm and proposed Hybrid CNN-SVM algorithm are displayed and they are compared.

7. Comparison Graph:

The displayed accuracies are compared in this stage.

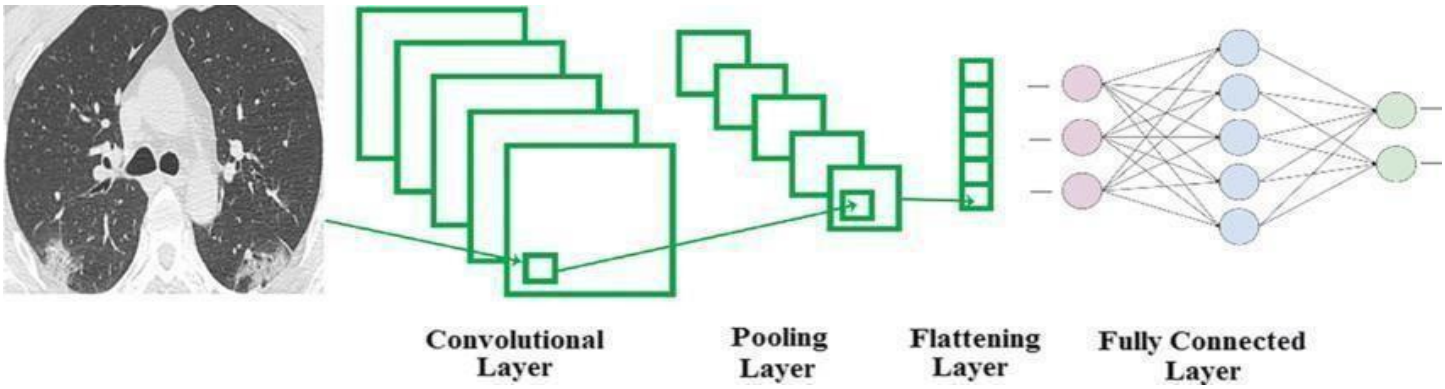


Fig 3.1 Architecture of CNN

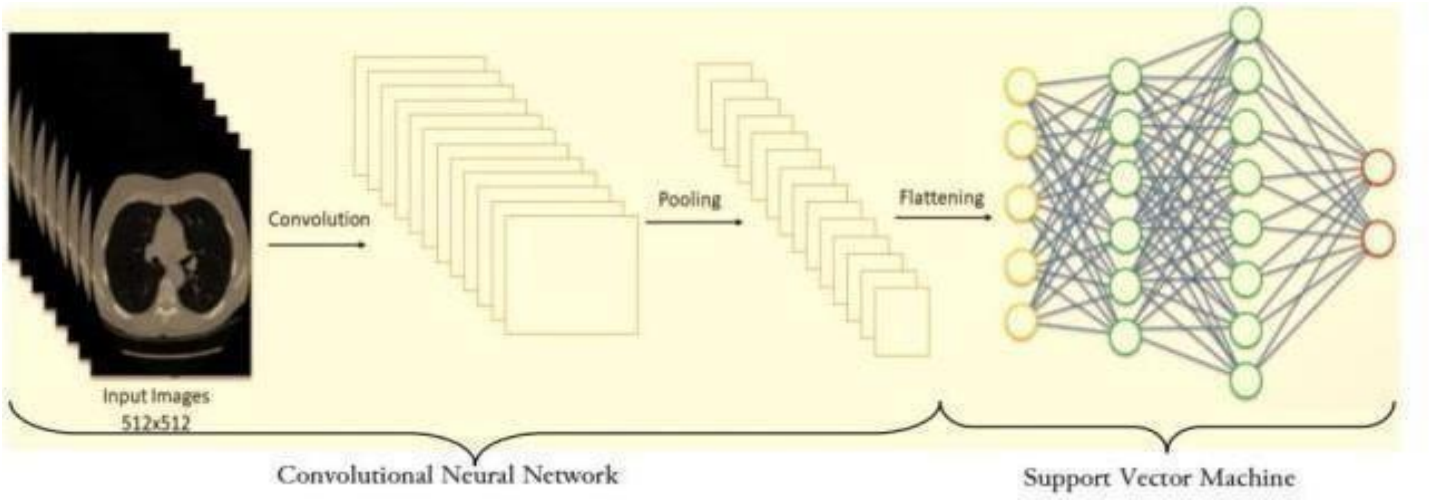


Fig 3.2 System Architecture

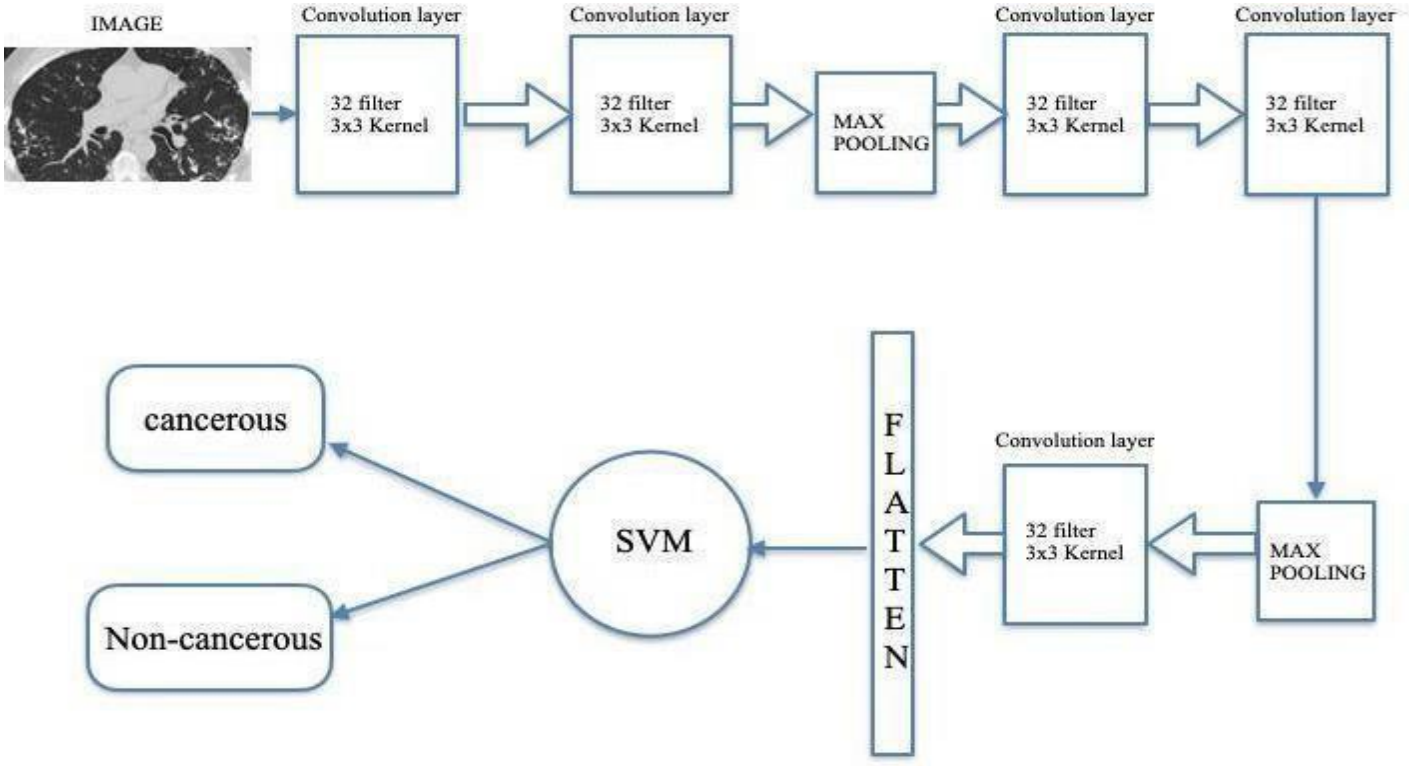


Fig 3.3 Block Diagram

3.3.2 Step by step explanation of Block Diagram:

Here's an explanation of how a dataset of lung images is processed in a hybrid CNN-SVM model for lung cancer detection:

1. Pre-processing:

- Data Acquisition: The process starts with collecting lung images, typically CT scans. These images might come from various sources like hospitals or public datasets [e.g., Lung Image Database Consortium (LIDC) dataset].
- Formatting: The images are then ensured to be in a consistent format (e.g., size, pixel depth) suitable for the model.
- Normalization: Normalization techniques might be applied to adjust the intensity values of the images for better training [e.g., scaling pixel values to a specific range like 0-1].

2. CNN Feature Extraction:

- Convolutional Layers: The pre-processed images are fed into the Convolutional Neural Network (CNN)
- portion of the model. Here, the CNN performs convolutions using filters to automatically extract features

from the images. These features capture low-level details like edges, textures, and patterns relevant for lung cancer detection.

- Pooling Layers: Pooling layers are often used alongside convolutional layers to reduce the dimensionality of the data and make the model more efficient. Techniques like max pooling select the maximum value from a local region, summarizing the activation of the previous layer.
- Multiple Stages: The CNN typically consists of multiple convolutional and pooling layers stacked together. Each layer learns progressively more complex features from the previous layer's output.

3. SVM Classification:

- Feature Vector: After feature extraction by the CNN, the resulting data is transformed into a feature vector, essentially a compressed representation containing the most relevant information from the images.
- SVM Training: This feature vector is then used to train a Support Vector Machine (SVM) classifier. During training, the SVM learns to distinguish between features representing cancerous and non-cancerous lung tissues based on labelled data provided.
- Classification: Once trained, the SVM can classify new lung images based on the extracted features. The model outputs a prediction, indicating whether the image is likely cancerous or not.

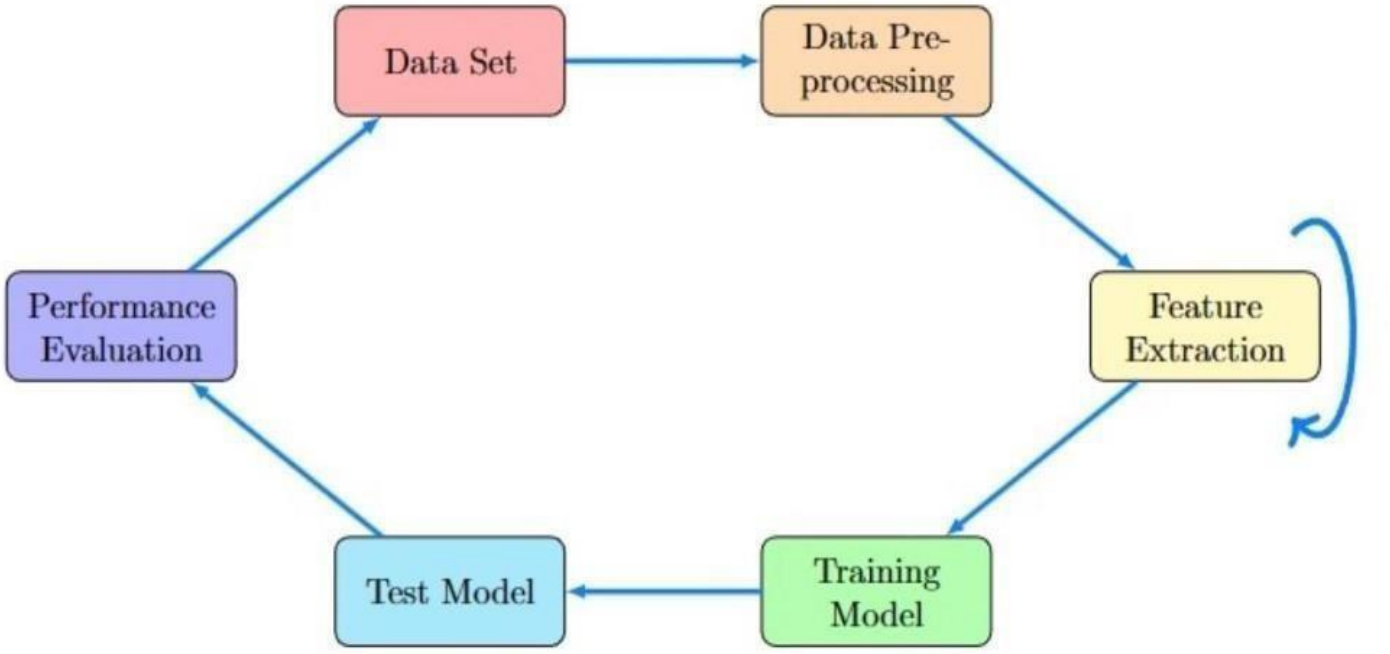


Fig 3.4 UML Diagram

The accomplishment of a hybrid CNN-SVM algorithm for lung cancer detection involves a mixture of both Convolutional Neural Networks (CNNs) and Support Vector Machines (SVM).

CHAPTER 4
RESULTS AND DISCUSSION

RESULTS AND DISCUSSION

4.1 Result

An experimental study was conducted on the suggested hybrid CNN-SVM model with the help of the lung image dataset obtained from Kaggle. This test scenario contains the images of lungs. These lung images are sent upon request. Diagnostic rules are then created from these images and transferred to the (SVM) for learning process. The learning process will start from its own process and finally it will check the images of the lungs for cancer. The last one is to evaluate whether the proposed method increases the accuracy of the detection. Accuracy refers to the predictions which are correct and is divided by the predictions of all the possible cases. The accuracy of the output is the key in determining the finest algorithm for future use. The more accurate algorithm gives the best output.

The Accuracy is calculated by using the formula:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} = \text{Number of Correct Predictions/ Total Number of Predictions}$$

where TP refers to True Positives, TN refers to True Negative, FP refers to the False Positive, and FN refers to False Negative.

The results we get will be in the form of GUI as shown in (Fig 4.1), the first stage is uploading dataset which we have downloaded from Kaggle which consists of both normal and abnormal lung images.

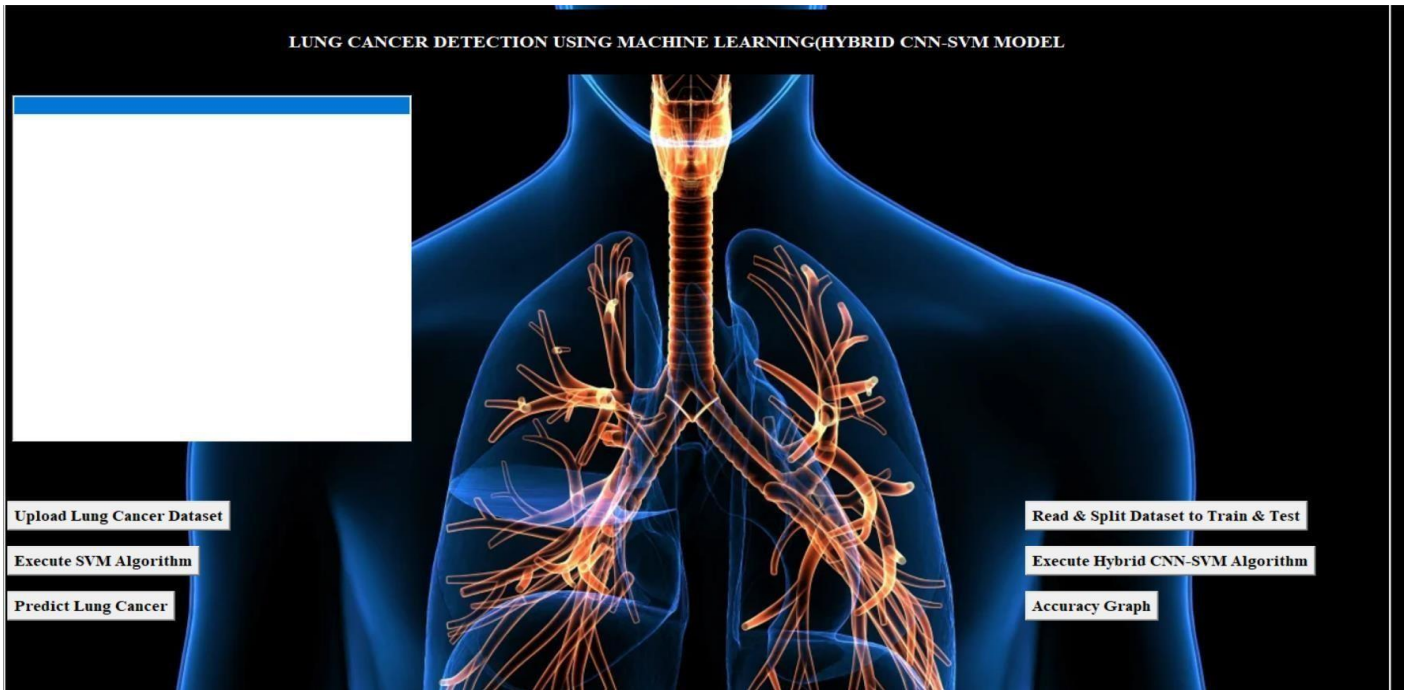


Fig 4.1 Interface after Executing the code

Steps to create GUI:

Tkinter is a built-in Python library that allows you to create graphical user interfaces (GUIs) for your applications.

1. Importing Tkinter:

You start by importing the tkinter module, often abbreviated as tk. This gives you access to all the necessary elements for building your GUI.

2. Creating the Main Window:

The main window serves as the foundation of your application. You use the tk.Tk() function to create this window object. This window will hold all the other GUI elements you add.

3. Adding Widgets:

Tkinter provides various widgets, which are the building blocks of your GUI. These widgets can be buttons, labels, text boxes, checkboxes, and more. You use specific functions for each widget type to create them and customize their properties like text, size, and position.

4. Layout Management:

Tkinter offers different layout managers to arrange your widgets within the main window. Common layout managers include:

Pack: Organizes widgets in a top-down or left-to-right manner, similar to packing boxes in a truck.

Grid: Places widgets in a grid-like structure, offering more precise control over their location.

Place: Allows you to specify the exact coordinates of each widget for absolute positioning.

5. Event Handling:

To make your GUI interactive, you define event handlers. These handlers specify what actions should occur when a user interacts with a widget, such as clicking a button or entering text.

6. Main Event Loop:

Finally, you use tk.mainloop() to start the main event loop. This loop keeps the GUI application running and continuously listens for user interactions and updates the window accordingly.

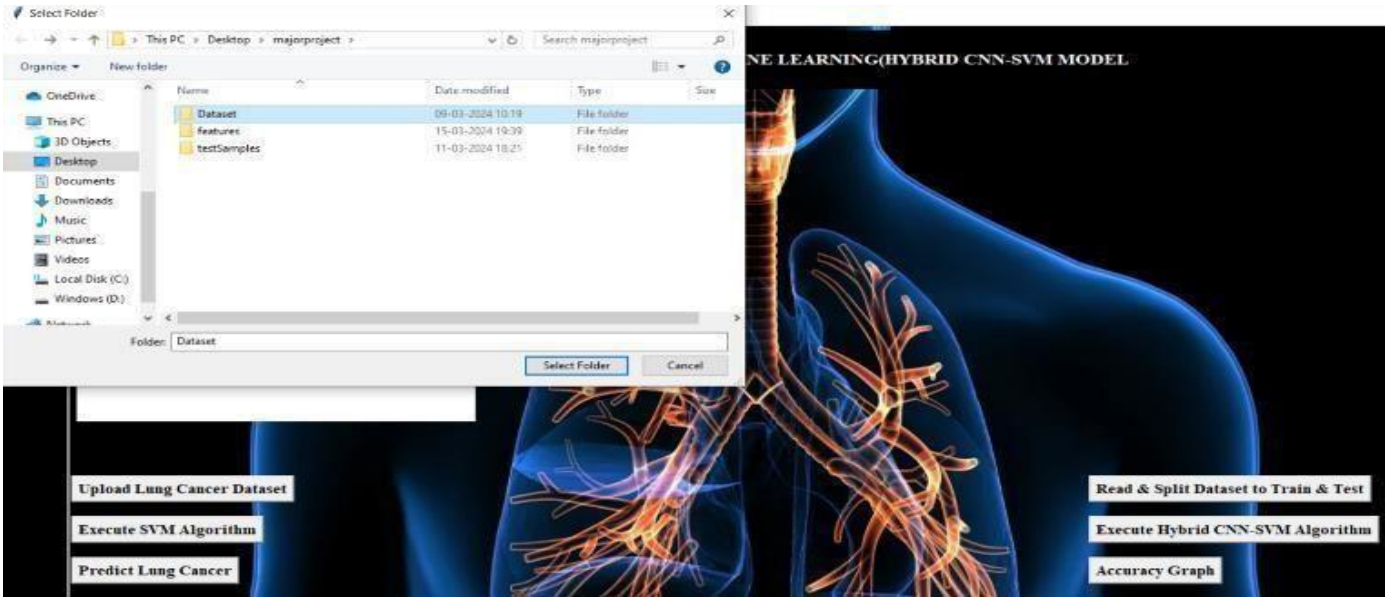


Fig 4.2 Interface to Upload Dataset

• After executing the code, we get the output in the form of Graphical user interface, in which it consists the button called Upload Lung Cancer Dataset, On clicking on the button we get the access of the file from where we can upload the dataset.

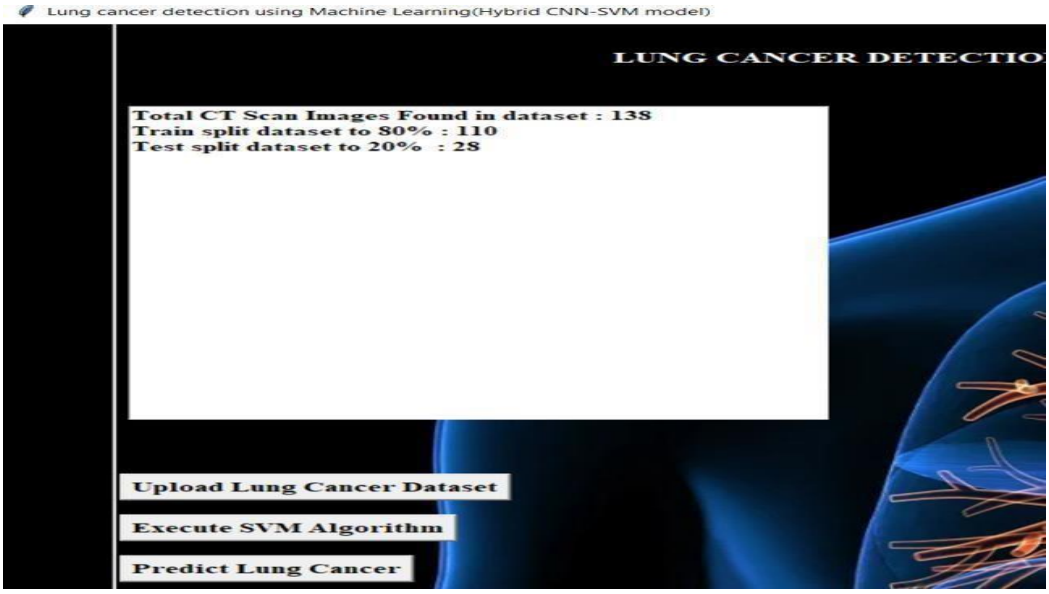


Fig 4.3 Splitting the dataset

• After uploading the Datasets, Split the dataset into training and testing sets that is 80% training, 20% testing.

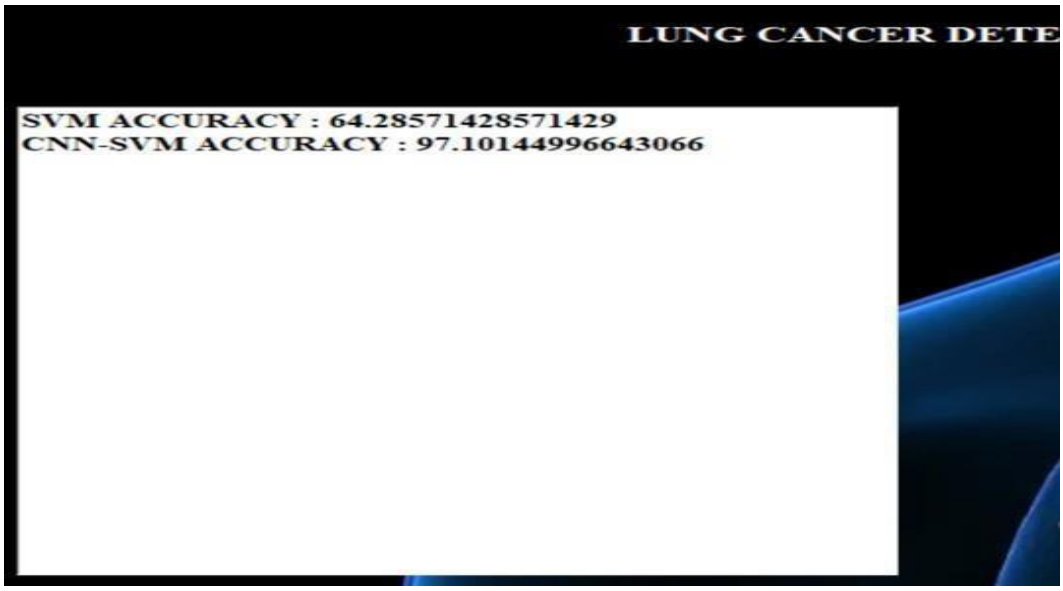


Fig 4.4 Accuracies of SVM alone and Hybrid CNN-SVM

• Next step is Executing normal SVM and Hybrid CNN-SVM Algorithm. In this stage CNN is executed using keras. Keras is a powerful library in python which is used to build the layers of CNN and in this process keras need not be installed separately because keras come along with the TensorFlow library. Here l2 regularization is used which is most frequently used in deep neural networks. It can be easily implemented using built-in functionality provided by keras application program interface.

• L2 regularization in deep learning, the regularization term is the sum of the squared values of all the weights in the neural network. It penalizes large weight values and encourages smaller weights, preventing any one weight from dominating the model.

• By comparing the accuracies of both normal SVM and Hybrid CNN-SVM model, the hybrid model is more accurate because it combines the strength of both CNN and SVM, here feature extraction power of CNN and classification power of SVM leads to generate higher accuracy compared to execution of individual algorithm.

Fig 4.5 Detecting the Test samples (Normal)

- Uploading the image and verifying whether it is normal or malignant.

- On Uploading the image dataset from the test sample folder, we get the out put as “imageis Normal(benign)” if the image doesn’t have any nodules or traces of carcinoma.

- The output for a normal lung image in a lung cancer detection system using a hybrid CNN-SVM model would depend on how the model is trained and what type of output it provides.

- Typically, in a binary classification scenario where the model aims to differentiate between normal and abnormal lung images (where abnormal includes lung cancer), the output could be in the form of a probability score or a binary label indicating the predicted class.

Fig 4.6 Detecting the Test samples (Abnormal)

- The model may output a probability score indicating the likelihood of the image containing lung cancer. A score close to 0 would indicate a low likelihood (normal lung), while a score close to 1 would indicate abnormal.

- In a binary classification setup, a binary label of 1 would typically indicate an abnormal or cancerous lung image. So, if the model outputs a binary label of 1 for a particular image, it suggests that the model predicts the presence of cancer in that lung image or else it is said to normal or benign image.

- After executing Hybrid CNN-SVM we get the output as the console shown in (Fig 4.7)

- In below screen you can see for CNN we use multiple filters to filter dataset for better prediction result and in above screen in first layer CNN use 62 X 62 image size with 32 filters and in second layer for 31 X 31 image size also it uses 32 filters and for each filter we will have best image features and prediction accuracy will be better.

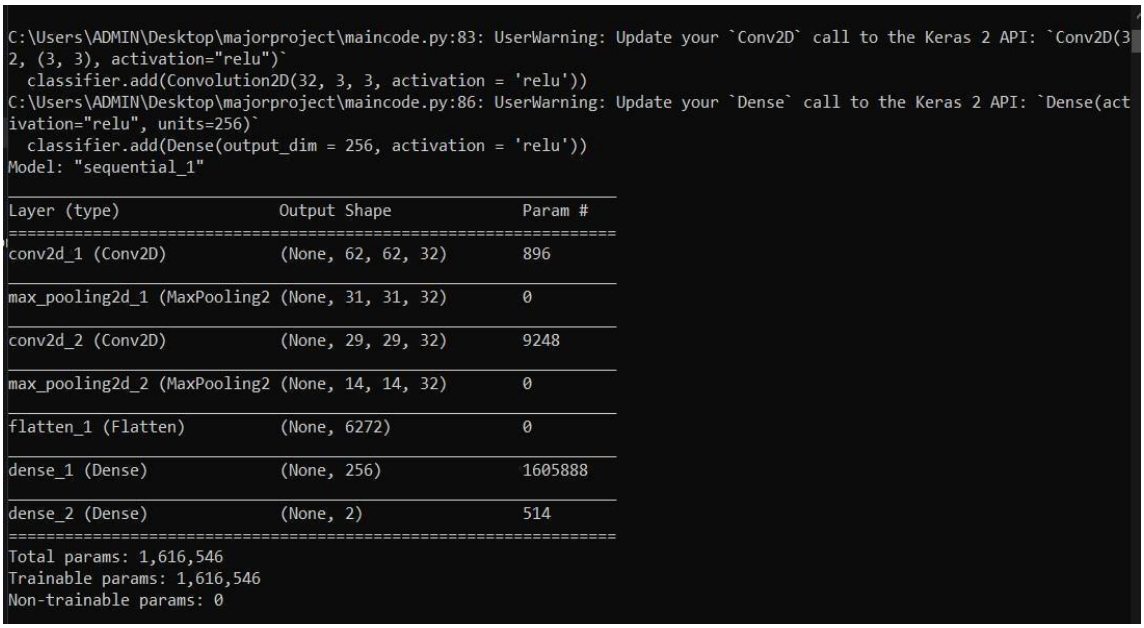


Fig 4.7 Console output after executing Hybrid CNN-SVM

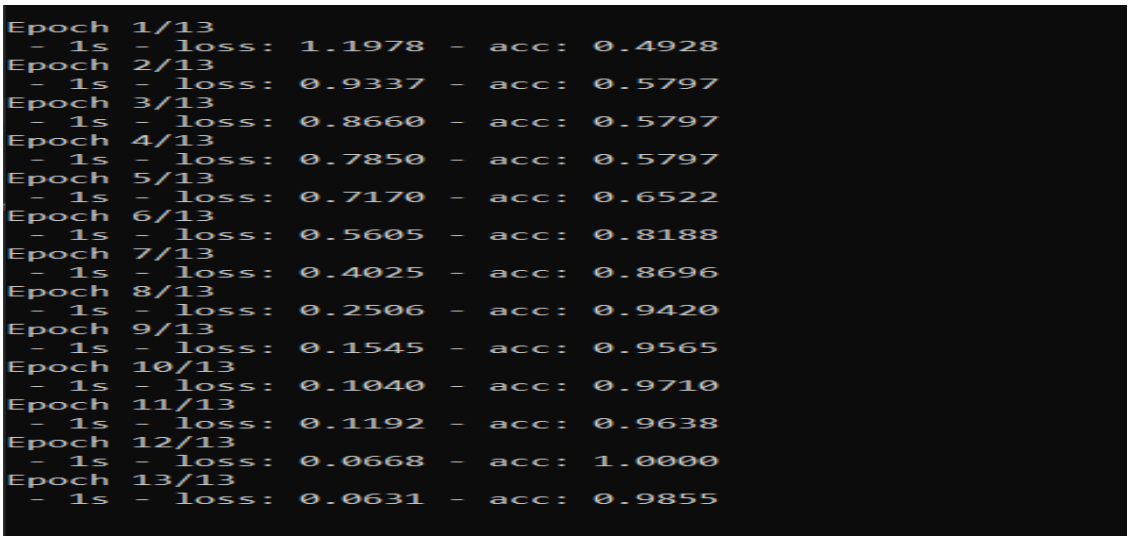


Fig 4.8 Console output after executing Hybrid CNN-SVM

- In above screen to run CNN we used 13 epoch/iteration and for each increase iteration accuracy get better and better and for last epoch we got 0.98% accuracy or even more or slightly less and the loss function decreases as the epoch gets executed continuously hence, we can say that the increasing of

- Accuracy and decreasing of loss function are positive indicators of the model's learning progress during training.

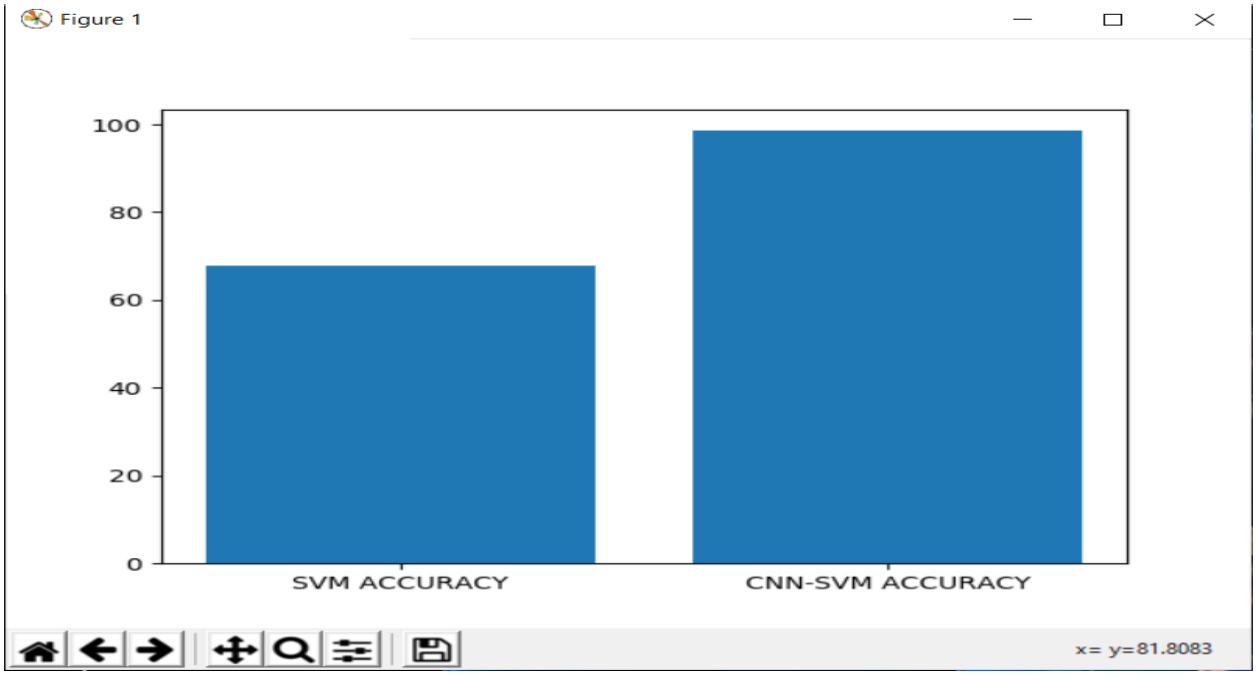


Fig 4.9 Comparison of Accuracy Graph

- The Accuracy of Ideal SVM or SVM alone is 64.2% and Hybrid CNN-SVM is 98.5% and here this is shown in terms of bar graph representation by using the python library called matplotlib.

- This is because the Hybrid CNN-SVM combines both the powers together that is the feature extraction power of CNN and binary classification power of SVM so due to this its giving high accuracy comparing to stand alone working algorithm like SVM.

CHAPTER 5
CONCLUSION

5.1 CONCLUSION

In summary, this study merges both CNN and SVM to detect tumor nodes including large cell carcinoma, adenocarcinoma, normal or squamous cell carcinoma on lung Computed Tomography images and verify if it is cancerous that is (Abnormal) or non-cancerous that is (Normal). The purpose of this study is to get a high stage of accuracy, which is the focus of all computer-assisted analysis. The methodology was implemented to all the chest X-ray image dataset, a standard and openly available set of X-ray lung images. The level of accuracy can be further elevated by increasing the set of the number of images used for the procedure. In addition, many types of x-rays, can be evaluated using this methodology. It should be possible to check all these pictures. By learning and analyzing the predictive results of various types of images, medical professionals will be able to use the most appropriate pictures to diagnose lung disease.

5.2 FUTURE WORK

In future we can Increase the diversity and quantity of the dataset by augmenting existing data with transformations such as rotation, scaling, flipping, and translation. This helps in making the model more robust. using different SVM kernels: Experiment with different kernel functions in SVM (e.g., linear, polynomial, Gaussian RBF) and kernel parameters to find the best combination that maximizes classification. Lastly, focusing on data quality, advanced CNN architectures, and domain-specific regularization techniques can be used. .

CHAPTER 6
REFERENCES

6.1 References

[1] https://www.nccn.org/professionals/physician_gls/default.aspx. Accessed Jan. 13, 2020.

[2] Gindi,A. M., Al Attiatalla, T. A., & Sami, M.M. (2014) "A Comparative Study for Comparing Two Feature Extraction Methods and Two Classifiers in Classification of Earlystage Lung CancerDiagnosis of chest x-ray images." Journal of American Science, 10(6): 13-22.

[3] Suzuki, K., Kusumoto, M., Watanabe, S. I., Tsuchiya, R., & Asamura, H. (2006) "Radiologic classification of small adenocarcinoma of the lung: radiologic-pathologic correlation and its prognostic impact," The Annals of Thoracic Surgery. 81(2): 413-419.

[4] Xiuhua,G., Tao, S., & Zhigang, L.(2011) "Prediction Models for Malignant Pulmonary Nodules Based-on Texture Features of CT Image." In Theory and Applications of CT Imaging and Analysis. DOI: 10.5772/14766.

[5] Mokhled S. Al-Tarawneh(August,2012),Lung Cancer Detection Using Image Processing Techniques.

[6] Non-Small Cell Lung Cancer, Available at: http://www.katemacintyrefoundation.org/ pdf/non- small-cell.pdf, Adapted from National Cancer Institute (NCI) and Patients Living with Cancer (PLWC), 2007, (accessed July 2011).

[7] Aggarwal, T., Furqan, A., & Kalra, K. (2015) "Feature extraction and LDA based classification of lung nodules in chest CT scan images." 2015 International Conference On Advances In Computing, Communications And Informatics (ICACCI), DOI: 10.1109/ICACCI.2015.7275773.

[8] Jin, X., Zhang, Y., & Jin, Q. (2016) "Pulmonary Nodule Detection Based on CT Images Using Convolution Neural Network." 2016 9Th International Symposium On Computational Intelligence And Design (ISCID). DOI: 10.1109/ISCID.2016.1053.

[9] Sangamithraa, P., & Govindaraju, S. (2016) "Lung tumour detection and classification using EK- Mean clustering." 2016 International Conference On Wireless Communications, Signal Processing And Networking (Wispnet). DOI: 10.1109/WiSPNET.2016.7566533.

[10] Roy, T., Sirohi, N., & Patle, A. (2015) "Classification of lung image and nodule detection using fuzzy inference system." International Conference On Computing, Communication & Automation. DOI: 10.1109/CCAA.2015.7148560.

[11] Ignatious, S., & Joseph, R. (2015) "Computer aided lung cancer detection system." 2015 Global Conference On Communication Technologies (GCCT), DOI: 10.1109/GCCT.2015.7342723.

[12] T. Saba Automated lung nodule detection and classification based on multiple classifiers voting Microsc Res Tech, 2019,pp. 1-9 10/1002/jemt.23326

[13] M. Firmino, G. Angelo, H. Morais, M.R. Dantas, R. Valentim Computer-aided detection (CADe) and diagnosis (CADx) system for lung cancer with likelihood of malignancy Biomed EngOnline, 15 (2016), p. 2.

[14] S.M. Naqi, M. Sharif, A. Jaffar Lung nodule detection and classification based on geometric fit in parametric form and deep learning Neural Comput Appl (2018), pp. 1-19

[15] Asuntha, A. Srinivasan Deep learning for lung Cancer detection and classification Multimed Tools Appl (2020), pp. 1-32

[16] S.A. Khan, M. Nazir, M.A. Khan, T. Saba, K. Javed, A. Rehman, *et al.* Lungs nodule detection framework from computed tomography images using support vector machine Microsc Res Tech (2019), 10.1002/jemt.23275.

[17] D. Kumar, A. Wong, D.A. Clausi Lung nodule classification using deep features in CT images2015 12th Conference on Computer and Robot Vision, IEEE (2015), pp. 133-138.

[18] M. Muzammil, I. Ali, I. U. Haq, A. A. Khaliq and S. Abdullah, "Pulmonary nodule classification using feature and ensemble learning-based fusion techniques," IEEE Access, vol. 9, pp. 113415–113427, 2021.

[19] G. Bansal, V. Chamola, P. Narang, S. Kumar and S. Raman, "Deep3Dscan: Deep residual network and morphological descriptor based framework for lung cancer classification and 3Dsegmentation,"IET Image Process, vol. 14, no. 7, pp. 1316–1326, 2020.

[20] G. Shah, R. Thammasudjarit, A. Thakkinstian and T.Suwatanapongched, "Nodulenet: A lung nodule classification using deep learning," Ramathibodi Medical Journal, vol. 43, no. 4, pp. 11– 19, 2020.

[21] Z. Guo, L. Xu, Y. Si and N. Razmjooy, "Novel computer-aided lung cancer detection based on convolutional neural network-based and feature-based classifiers using metaheuristics," International Journal of Imaging System and Technology, vol. 31, no. 4, pp. 1954–1969, 2021.

[22] Kanchan Pradhan & Priyanka Chawla. "Medical Internet of Things using Machine Learning for Lung cancer detection", Journal of Management Analytics, vol 7, no. 4, published on- 2020.

[23] Rajini A, Jabbar M.A."Lung cancer prediction using Random Forest", Bentham Science Publishers, vol 14, no 5, pp 1650-1657, 2021.

[24] www.kaggle.com

CHAPTER 7
APPENDIX

7.1 Code

```
from tkinter import messagebox
from tkinter import *
from tkinter import simpledialog
import tkinter
from tkinter import filedialog
import matplotlib.pyplot as plt
import numpy as np
from tkinter.filedialog import askopenfilename
import tensorflow as tf
import pandas as pd
import os
import cv2
import numpy as np
from sklearn import svm
from sklearn.metrics import accuracy_score
from sklearn.model_selection import train_test_split
from sklearn.decomposition import PCA
from keras.utils.np_utils import to_categorical
from keras.layers import MaxPooling2D
from keras.layers import Dense, Dropout, Activation, Flatten
from keras.layers import Convolution2D
from keras.models import Sequential
from PIL import ImageTk, Image
from tensorflow.keras.regularizers import l2
#from tensorflow.keras.models import Sequential


main = tkinter.Tk()
main.title("Lung cancer detection using Machine Learning(Hybrid CNN-SVM model)")
main.geometry("1300x1200")

global filename
global classifier
global svm_sr, cnn_sr
global X, Y
global X_train, X_test, y_train, y_test
global pca

def uploadDataset():
    global filename
    filename = filedialog.askdirectory(initialdir=".")
    text.delete('1.0', END)
    text.insert(END,filename+" loaded\n");
```

```
def splitDataset():
    global X, Y
    global X_train, X_test, y_train, y_test
    global pca
    text.delete('1.0', END)
    X = np.load('features/X.txt.npy')
    Y = np.load('features/Y.txt.npy')
    X = np.reshape(X, (X.shape[0],(X.shape[1]*X.shape[2]*X.shape[3])))

    pca = PCA(n_components = 100)
    X = pca.fit_transform(X)
    print(X.shape)
    X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.2)
    text.insert(END,"Total CT Scan Images Found in dataset : "+str(len(X))+"\n")
    text.insert(END,"Train split dataset to 80% : "+str(len(X_train))+"\n")
    text.insert(END,"Test split dataset to 20% : "+str(len(X_test))+"\n")


def  executeSVM():
    global   classifier
    global svm_sr
    text.delete('1.0', END)
    cls = svm.SVC()
    cls.fit(X_train, y_train)
    predict = cls.predict(X_test)
    svm_sr = accuracy_score(y_test,predict) * 100
    classifier = cls
    text.insert(END,"SVM ACCURACY : "+str(svm_sr)+"\n")

def  executeHybridmodel():
    global cnn_sr
    X = np.load('features/X.txt.npy')
    Y = np.load('features/Y.txt.npy')
    Y = to_categorical(Y)
    classifier = Sequential()
    classifier.add(Convolution2D(32, 3, 3, input_shape = (64, 64, 3), activation = 'relu'))
    classifier.add(MaxPooling2D(pool_size = (2, 2)))
    classifier.add(Convolution2D(32, 3, 3, activation = 'relu'))
    classifier.add(MaxPooling2D(pool_size = (2, 2)))
    classifier.add(Flatten())
    classifier.add(Dense(output_dim = 256, activation = 'relu'))
    classifier.add(Dense(2, kernel_regularizer=tf.keras.regularizers.l2(0.01),activation ='linear'))
    print(classifier.summary())
    classifier.compile(optimizer = 'adam', loss = 'hinge', metrics = ['acc'])
    hist = classifier.fit(X, Y, batch_size=16, epochs=13, shuffle=True, verbose=2)
```

```
  hist = hist.history
    acc = hist['acc']
    cnn_sr = acc[12] * 100
    text.insert(END,"CNN-SVM ACCURACY : "+str(cnn_sr)+"\n")

def predictCancer():
    filename = filedialog.askopenfilename(initialdir="testSamples")
    img = cv2.imread(filename)
    img = cv2.resize(img, (64,64))
    im2arr = np.array(img)
    im2arr = im2arr.reshape(64,64,3)
    im2arr = im2arr.astype('float32')
    im2arr = im2arr/255
    test = []
    test.append(im2arr)
    test = np.asarray(test)
    test = np.reshape(test, (test.shape[0],(test.shape[1]*test.shape[2]*test.shape[3])))
    test = pca.transform(test)
    predict = classifier.predict(test)[0]
    msg = ''
    if predict == 0:
        msg = "image is Normal(benign)"
    if predict == 1:
        msg = "image is Abnormal(malignant)"
    img = cv2.imread(filename)
    img = cv2.resize(img, (400,400))
    cv2.putText(img, msg, (10, 25), cv2.FONT_HERSHEY_SIMPLEX,0.7, (0, 255, 255), 2)
    cv2.imshow(msg, img)
    cv2.waitKey(0)

def graph():
    height = [svm_sr, cnn_sr]
    bars = ('SVM ACCURACY','CNN-SVM ACCURACY')
    y_pos = np.arange(len(bars))
    plt.bar(y_pos, height)
    plt.xticks(y_pos, bars)
    plt.show()

width= main.winfo_screenwidth()
height= main.winfo_screenheight()

main.geometry("%dx%d" % (width, height))
main.resizable(False, False)
```

```
imgTemp = Image.open("C:/Users/ADMIN/Desktop/majorproject/lung2.webp")
img2 = imgTemp.resize((1400,900))
img = ImageTk.PhotoImage(img2)


label = Label(main,image=img)
label.pack(side='top',fill=Y,expand=True)

font = ('times', 14, 'bold')
title = Label(main, text='LUNG CANCER DETECTION USING MACHINE LEARNING(HYBRID
CNN-SVM MODEL)')
title.config(bg='Black', fg='white')
title.config(font=font)
title.config(height=3, width=120)
title.place(x=85,y=5)

font1 = ('times', 12, 'bold')
text=Text(main,height=20,width=50)
scroll=Scrollbar(text)
text.configure(yscrollcommand=scroll.set)
text.place(x=75,y=100)
text.config(font=font1)


font1 = ('times', 13, 'bold')
uploadButton = Button(main, text="Upload Lung Cancer Dataset", command=uploadDataset)
uploadButton.place(x=70,y=550)
uploadButton.config(font=font1)

readButton = Button(main, text="Read & Split Dataset to Train & Test", command=splitDataset)
readButton.place(x=1100,y=550)
readButton.config(font=font1)

svmButton = Button(main, text="Execute SVM Algorithm", command=executeSVM)
svmButton.place(x=70,y=600)
svmButton.config(font=font1)

kmeansButton = Button(main, text="Execute Hybrid CNN-SVM Algorithm", command=executeHybridmodel)
kmeansButton.place(x=1100,y=600)
kmeansButton.config(font=font1)

predictButton = Button(main, text="Predict Lung Cancer", command=predictCancer)
predictButton.place(x=70,y=650)
predictButton.config(font=font1)

graphButton = Button(main, text="Accuracy Graph", command=graph)
graphButton.place(x=1100,y=650)
graphButton.config(font=font1)
```

```
main.config(bg='Black')
main.mainloop()
```

8 GITHUB LINK

https://github.com/Ganesh-2402/Lung-cancer-detection-using-ml

9 DOI

https://doi.org/10.22214/ijraset.2024.59257

PUBLISHED PAPER

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue III Mar 2024- Available at www.ijraset.com

Lung Cancer Detection Using Machine Learning

Mr. D. Narsimha Reddy[1], C. Ganesh[2], K. Shiva Abhigna[3], P. Tharun Sai[4]
[1]*Associate Professor,* [2, 3, 4] *UG Student, Department of CSE, CMR College of Engineering & Technology, Hyderabad, Telangana*

Abstract: ***Lung cancer ranks among the primary causes of death on global scale. Catching this disease early can increase your chances or opportunities of survival. Computer-assisted detection (CAD) is used to create CT images and even X-rays of the lungs to determine whether cancer is present in the images. This paper represents an image classification by the combination of a neural network (CNN) algorithm and support vector machine (SVM). The algorithm spontaneously separates and analyzes lung picture or image to detect cancer cells. Compared to full-scale networks, CNNs are easier to train and have less overhead. We introduce CNN-SVM because it has accurate performance than other existing terminologies. The merits of this method are that it can detect cancer on the CT image***

Keywords: CNN, Classification , SVM, CAD, Lung cancer

I. INTRODUCTION

Pulmonary carcinoma screening is the process of determining the appearance or occurrence of lung disease or cancer in a person. This process can be done through a variety of medical technologies and procedures, including diagnostic tools and even machine learning and deep learning. Lung cancer generally falls into two primary categories: small cell lung carcinoma and non-small cell lung carcinoma (adenocarcinoma and squamous cell carcinoma are subtypes). These distinct types of lung carcinoma have different growth and also, they are differently treated. Non- small cell lung cancer prevails more in number than small cell lung cancer [1].

Lung carcinoma is one of the types of cancer. It is highly difficult to diagnose because it occurs in the last stage and shows symptoms. However, rate of mortality and morbidity can be decreased with early diagnosis of the disease. CT imaging, the best imaging technology, is reliable for lung cancer diagnosis because it can reveal any desired and invisible lung cancer lump [2]. However, differences between computed tomography images and the determination of anatomical structures by the physicians and the radiologists may cause diagnostic problems in the collection of cancer cells [3].

In recent years, computerized diagnostics has emerged as an additional and promising tool to assist radiologists and physicians in diagnosing cancer [4]. Many systems and studies have been developed for the identification or recognition of lung cancer through various examinations as well as analysis.

However, the detection accuracy of some systems is not satisfactory, and some systems still need to be gained to achieve the highest accuracy close to 100%. Machine learning and Image processing are used in the early detection and classification process of lung cancer.

II. RELATED WORK

In this section we have studied various implementations of lung carcinoma detections & we summarized our findings and concluded by researching & referencing various papers. They are: Mokhled S. Al-Tarawneh [5]. Lung cancer is a deadly disease in which abnormal cells multiply and develop into tumors. Cancer cells can be removed from the lungs by the blood or lymphatic fluid around the lungs. Lymph flows through lymphatic vessels to the lymph nodes in the middle of the lungs and chest. Lung cancer often spreads to the chest area because the lymph nodes in the lung drain into the area of chest. Metastasis comes when cancer lumps or cells exit their site of origin and migrate via blood vessels to the tumor or other parts of body [6]. Many researchers have proposed and applied different imaging techniques and machine learning to diagnose lung cancer.

Aggarwal, Furquan and Kalra [7] proposed a system which allows the process of classification of nodules and lung organs. This method pull-out statistical, geometric and grayscale features. LDA is used as the best method for classification and segmentation. The system provides accuracy of 84%, sensitivity of 97.14%, and the specificity of 53.33%. Although the system has detected cancer cells, its accuracy is not accepted. Machine learning techniques were not used for classification and simple segmentation methods were preferred. It is therefore not better to combine one of its steps in our new proposed model.

Jin, Zhang, and Jin [8] used the deep learning methodology called convolutional neural networks as classifiers in CAD systems to diagnose carcinoma of lungs. The system provides 84.6% of accuracy and 82.5% of sensitivity, and the specificity of 86.7%.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue III Mar 2024- Available at www.ijraset.com

The merit of this method is that it uses circular filters in the region of interest extraction stage, thus decreasing cost of the training and analysis steps. Even though its usage rate has decreased, its accuracy is still not good.

Sangamithraa and Govindaraju [9] used the ML algorithm that is K-means unsupervised learning algorithm for the process of clustering or segmentation. It forms the group of data of pixels on the basis of certain characteristics. The model uses a backpropagation network for classification. Entropy, correlation, uniformity, SSIM, PSNR etc. Use the Gray Level Co-occurrence Matrix method to extract features such as The accuracy of the system is approximately 90.7%. Image Preprocessing Median filters are used to remove noise; This is important for our new model to overcome from noise and improve the accuracy performance.

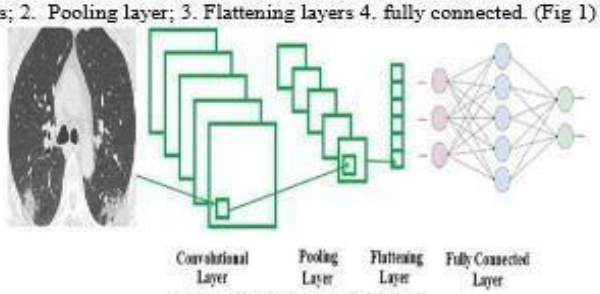
Roy, Sirohi and Patle [10] developed a system for the diagnosis of lung cancer nodules using a fuzzy relationship and a function model. This system uses grayscale conversion to improve contrasting of image. Image is binarized before segmentation and the image is segmented as a result using a reference contour model. Cancer is classified using a technique called fuzzy inference techniques. Area, mean, entropy, correlation, major axis length and minor axis length, etc. Remove features like. to show the classifier. Overall system accuracy is 94.12%. Given its limitations, this model does not classify future tumors as benign (cancerous) or malignant which is termed as (cancerous).

III. PROPOSED METHODOLOGY

Advance diagnosis of lung cancer can reduce mortality. It is main principal to catch cancer early, prevent its development, and eliminate it early before it starts to grow rapidly. Technology and deep learning are broadly used in medicine to monitor, detect, classify, and predict diseases. Our system includes a CNN-SVM architecture that removes and deletes redundant information that affects accuracy. This is done in the step of integrating the CNN architecture used to determine cancer cells in a (FCN) using the modified version of SVM model.

1) *Convolutional Neural Network (CNN):* The deep learning methodology preferred in this research-based study is a convolutional neural network (CNN). It is known for multilayer feedforward neural network which is biologically influenced [11]. CNN has many layers, and they can be separated in three stages: convolutional (which computes the output of regional connections in neurons), max pooling (subsampling of inputs), and fully connected layer (used to calculate the Activation of every class).

The input source of the CNN is an (a x a x b) image; where a is the width and height, b is the total number of channels, and in the convolutional process there will be k convolution filters of size n x n, where n < a. Create a CNN and fit the data. The CNN model has 4 steps: 1. Convolution layers; 2. Pooling layer; 3. Flattening layers 4. fully connected. (Fig 1) [12] (CNN architecture)



(fig-1 Architecture of CNN)

In the convolutional layer, feature maps are defined by dot product process of the input image and feature detectors. More feature maps are produced and the size of the image is reduced for uncomplicated processing. Since many detectors are used, different features have been developed. In this step, the ReLU Rectified linear unit activation function is used to add nonlinearity to CNN (since the image is generally nonlinear). In pooling layer stage, the resolution of feature maps is decreased; This step uses maximum pooling.

At this stage, a lot of unnecessary data is removed, which has a positive impact on getting good results because irrelevant/insignificant data will not be entered into a completely connected process. In the flattening stage, the feature map of pooled form which are the output of pooling layer is flattened into a single column or block. This is done to pass this block or column into the Support Vector Machine (SVM) model. Finally, all the features are processed in the merging process of SVMs, resulting in one of four options: cancers like adenocarcinoma, and greater cell carcinoma, normal(non-cancerous), or squamous cell carcinoma.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue III Mar 2024- Available at www.ijraset.com

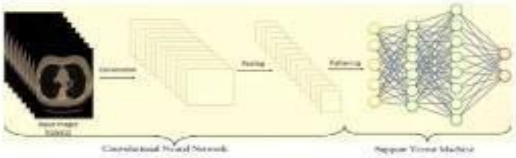
a) *Convolution layer:* Convolution layer is the important building block of CNN. It has layers (or kernels) that it has not learned throughout training. The size of the filter is generally smaller than the actual image. Each filter is equipped with images and creates activation map. For convolution, the filter slides across the height and width of the image, and the point features of each element of the filter and input are calculated at each position.

b) *Pooling layer:* Pooling layer is used to decrease the feature map. Therefore, it reduces the number of unlearned information and the computational cost in the network. This step shows the available features in the feature map created by the convolution layer. Therefore, additional processing is performed on content features rather than the fully localized features that convolutional layers do. This makes the model more robust to changes at certain locations in the input image. These are two types 1) max pooling 2) min pooling.

c) *Flattening layer:* Flattening is used to transform all two-dimensional arrays created by combining the image into a long chain (single long continuous linear vector). The flattened matrix is fed as input to all compositing operations to segment the image i.e fully connected layer.

d) *Fully connected layer:* this layer is feed forward neural network; these are the last few layers of CNN. The input to this layer is the output of pooling and flattening layers output.

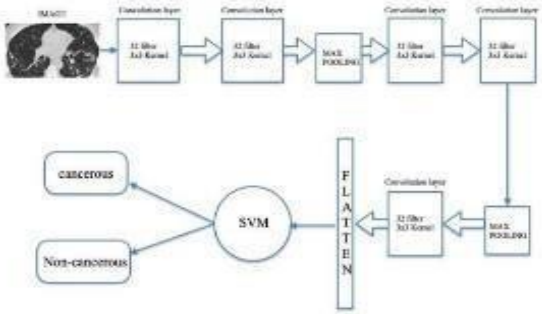
2) *Support Vector Machine (SVM):* SVM is one of the prominent machine learning algorithm used for learning techniques used in classification and also for regression problems. Even though, in machine learning concept it is widely used only in classification problems which classify two different classes. The motivation and purpose of the SVM algorithm is to create a boundary of decision which can divide the space of n-dimensional or spatial space into clusters so that we can place new data into correct clusters in future. This clear-cut boundary is called as hyperplane. SVM chooses point vectors which can help to creates a general plane. These states are called support vectors, so the algorithm is called as SVM.



(fig-2 Architecture of hybrid CNN-SVM)

A Hybrid CNN-SVM model merges both Convolutional Neural Networks (CNNs) and the ML algorithm as Support Vector Machines (SVM) algorithms. In this approach, in this model the last layer of CNN is replaced by SVM classifier, the CNN is typically used for feature extraction and representation learning from image data, while SVM is employed for classification. The CNN extracts relevant features from the input data, and these features are then fed into the SVM for making predictions. This hybrid model leverages the strengths of both CNNs, known for their effectiveness in learning hierarchical characteristics from raw data, and SVMs, known for their robustness in classification tasks.

IV. BLOCK DIAGRAM



(fig-3 Block Diagram of hybrid CNN-SVM)

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue III Mar 2024- Available at www.ijraset.com

V. IMPLEMENTATION

The accomplishment of a hybrid CNN-SVM algorithm for lung cancer detection involves a mixture of both Convolutional Neural Networks (CNNs) and Support Vector Machines (SVM). Where CNNs are used extraction of characteristics from images with the classification capabilities of SVM. It starts with the procedure of data collection by obtaining a dataset containing images of lung scans along with corresponding labels indicating whether each scan contains cancerous regions or not. It's crucial to have a diverse and representative dataset for training a robust model.

In addition, by following the method of preprocessing it preprocesses the images to standardize their size, resolution, and intensity levels. Common preprocessing steps include resizing, normalization to enhance the model's ability to discover the unseen data. Then it extracts the features by training a CNN architecture to draw out meaningful features from the lung scan images.

The CNN is typically composed of layers like convolutional for feature extraction, pooling layers and finally fully connected layers for classification. Then it represents the feature after training the CNN and uses it as a feature extractor to obtain feature representations (vectors) for each image in the dataset. The output of one of the last layers before the classification layer can be considered as high-level features representing the input image. Training SVM is one of the important steps using the extracted features as inputs to train an SVM classifier. SVMs are powerful discriminative models that aim to find the optimal hyperplane to separate different classes in the feature space. By training an SVM on the CNN-extracted features, we leverage the discriminative power of SVMs to perform the last classification task.

Assess the performance of the hybrid CNN-SVM model using numerous measures such as recall, precision, accuracy, F1- score, and area under the curve of ROC (AUC). Split the dataset into validation, training, and test sets to assess the model generalization ability. Once the hybrid model achieves satisfactory performance on the validation set, deploy it in a real-world setting for lung cancer detection. Continuously monitor and optimize the model's presentation over time, considering factors such as data drift, model drift, and emerging clinical insights. By combining the feature extraction capabilities of CNNs with the classification prowess of SVMs, the hybrid approach aims to leverage the complementary strengths of both models to enhance lung cancer detection accuracy and robustness.

VI. RESULTS AND DISCUSSIONS

An experimental study was conducted on the suggested hybrid CNN-SVM model with the help of the lung image dataset obtained from Kaggle. This test scenario contains the images of lungs. These lung images are sent upon request. Diagnostic rules are then created from these images and transferred to the (SVM) for learning process. The learning process will start from its own process and finally it will check the images of the lungs for cancer. The last one is to evaluate whether the proposed method increases the accuracy of the detection. Accuracy refers to the predictions which are correct and is divided by the predictions of all the possible cases. The accuracy of the output is the key in determining the finest algorithm for future use. The more accurate algorithm gives the best output.

The results we get will be in the form of GUI, the first stage is uploading dataset which we have downloaded from Kaggle which consists of both normal and abnormal lung images.

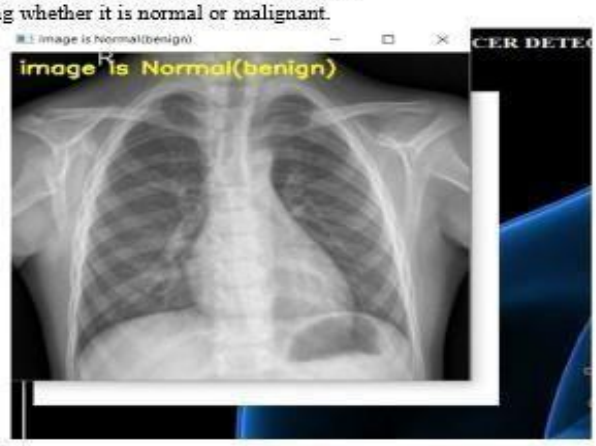
The second stage is splitting the dataset into train and test category

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue III Mar 2024- Available at www.ijraset.com

Next step is Executing normal SVM and Hybrid CNN-SVM Algorithm. In this stage CNN is executed using keras. Keras is a powerful library in python which is used to build the layers of CNN and in this process keras need not be installed separately because keras come along with the TensorFlow library. Here l2 regularization is used which is most frequently used in deep neural networks. It can be easily implemented using built-in functionality provided by keras application program interface.

By comparing the accuracies of both normal SVM and Hybrid CNN-SVM model, the hybrid model is more accurate because it combines the strength of both CNN and SVM, here feature extraction power of CNN and classification power of SVM leads to generate higher accuracy compared to execution of individual algorithm.

Uploading the image and verifying whether it is normal or malignant.



International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue III Mar 2024- Available at www.ijraset.com

Table-1. Comparison with published papers.

Sno	Published paper	Dataset using	Accuracy in %
1	CNN [13]	LIDC	79.40%
2	Reinforcement learning (ANN) [14]	LUNA	64.4%
3	DBNs [13]	LIDC	81.19%
4	SDAE [13]	LIDC	79.29%

VII. CONCLUSION

In summary, this study merges both CNN and SVM to detect tumor nodes including large cell carcinoma, adenocarcinoma, normal or squamous cell carcinoma on lung Computed Tomography images and verify if it is cancerous that is (Abnormal) or non-cancerous that is (Normal). The purpose of this study is to get a high stage of accuracy, which is the focus of all computer-assisted analysis. The methodology was implemented to all the chest CT scan image dataset, a standard and openly available set of CT images.

The level of accuracy can be further elevated by increasing the set of the number of images used for the procedure. In addition, many types of x-rays, can be evaluated using this methodology. It should be possible to check all these pictures. By learning and analyzing the predictive results of various types of images, medical professionals will be able to use the most appropriate pictures to diagnose lung disease.

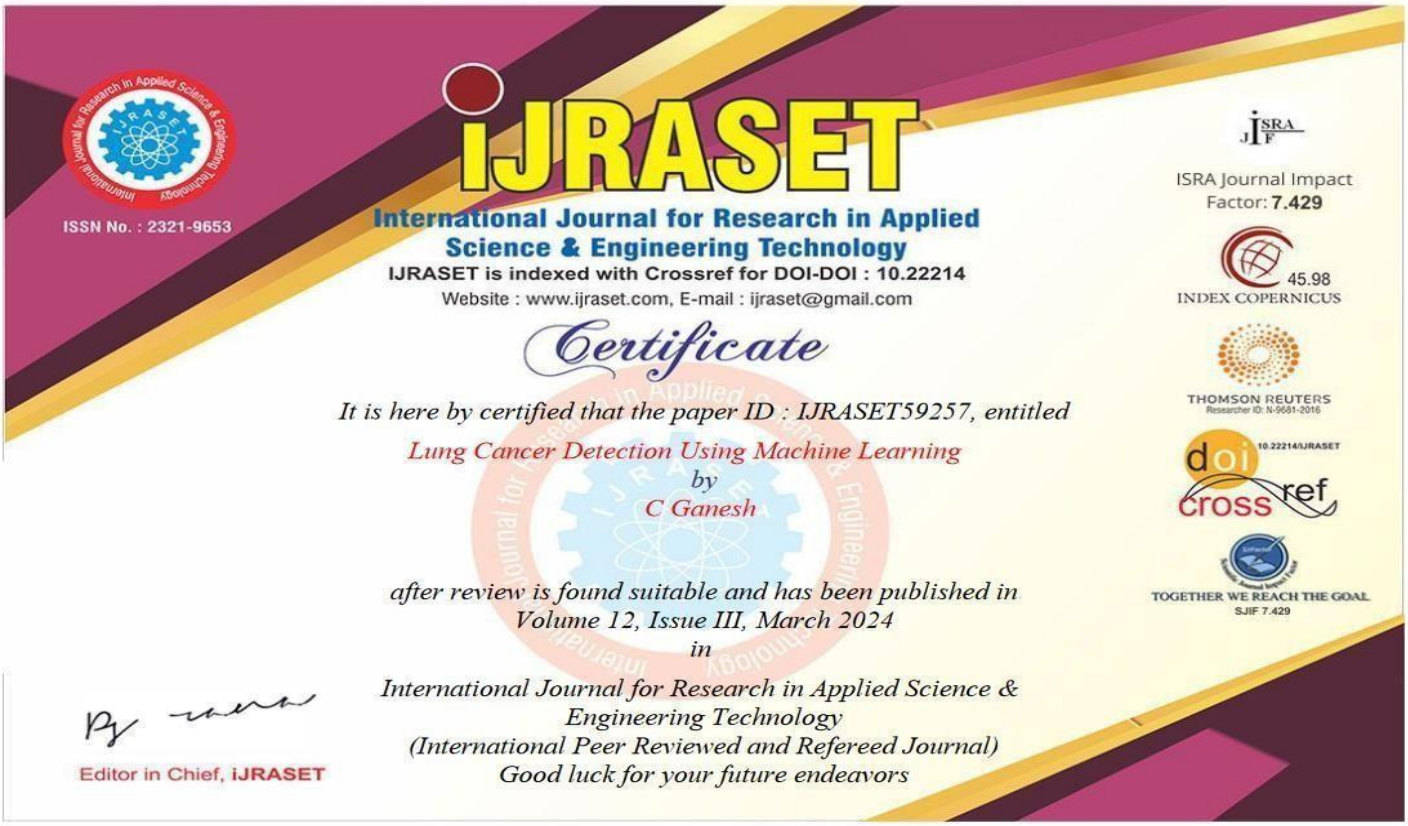
REFERENCES

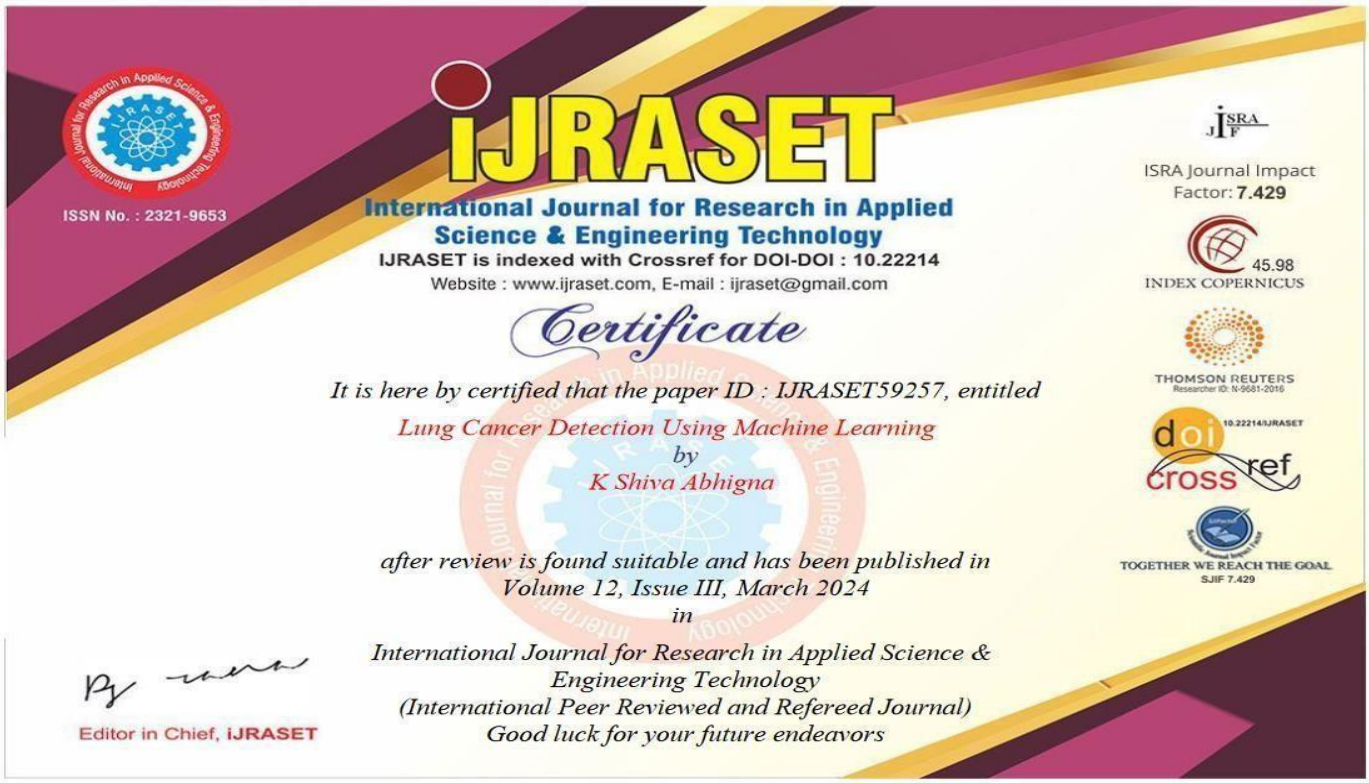
[1] https://www.nccn.org/professionals/physician_gls/default.aspx.
[2] Gindi,A. M., Al Attiatalla, T. A., & Sami, M.M. "A Comparative Study for Comparing Two Feature Extraction Methods and Two Classifiers in Classification of Earlystage Lung Cancer or disease Diagnosis of the chest x-ray images(2014)." Journal of the American Science, 10(6): 13-22.
[3] Suzuki, K., Kusumoto, M., Watanabe, S. I., Tsuchiya, R.,& Asamura, H. (2006) "Radiologic classification of small adenocarcinoma of the lung: radiologic-pathologic correlation and its prognostic impact," The Annals of Thoracic Surgery. 813-419.
[4] Xiuhua,G., Tao, S., & Zhigang, L. "Prediction Models for Malignant Pulmonary Nodules Based-on Texture Features of CT Image." In Theory and Applications of CT Imaging and Analysis. DOI: 10.5772/14766.
[5] Mokhled S. Al-Tarawneh(August,2012),Lung Cancer Detection Using Image Processing Techniques. K. Elissa, "Title of paper if known," unpublished.
[6] http://www.katemacintyrefoundation.org/pdf/nonsmall-cell.pdf, Adapted from National Cancer Institute (NCI) and Patients Living with Cancer (PLWC), 2007, (accessed July 2011).
[7] Aggarwal, T., Furqan, A., & Kalra, K. (2015) "Feature extraction and LDA based classification of lung nodules in chest CT scan images2015 (ICACCI), DOI: 10.1109/ICACCI.2015.7275773
[8] Jin, X., Zhang, Y., & Jin, Q. (2016) "Pulmonary Nodule Detection Based on CT Images Using Convolution Neural Network." 2016 9Th International Symposium On Computational Intelligence And Design (ISCID). DOI: 10.1109/ISCID.2016.1053.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue III Mar 2024- Available at www.ijraset.com

[9] Sangamithraa, P., & Govindaraju, S. (2016) "Lung tumour detection and classification using EKMean clustering." 2016 International Conference On Wireless Communications, Signal Processing And Networking (Wispnet). DOI: 10.1109/WiSPNET.2016.7566533.
[10] Roy, T., Sirohi, N., & Patle, A. (2015) "Classification of lung image and nodule detection using fuzzy inference system." International Conference On Communication & Automation. DOI: 10.1109/CCAA.2015.7148560.
[11] H.-y. Lee, "Deep learning tutorial," Open Course, Online Available, 2020. Available at: Google Scholar.
[12] https://www.researchgate.net/figure/General-architecture] X.-F. Cao, Y. Li, H.-N. Xin, H.-R. Zhang, M. Pai, and L. Gao,"Application of artificial intelligence in digital chest radiography reading for pulmonary tuberculosis screening," Chronic Diseases and Translational Medicine, vol. 7, no. 1, pp. 35-40, 2021. doi: 10.1016/j.cdtm.2021.02.001
[13] H. Wang et al., "Comparison of machine learning methods for classifying mediastinal lymph node metastasis of non-small cell lung cancer from 18 F-FDG PET/CT images," EJNMMI research, vol., no.1, pp. 1-11, 2017. doi: 10.1186/s13550-017-0260-9


doi
cross ref
10.22214/IJRASET
INDEX COPERNICUS
45.98
JISRA F
IMPACT FACTOR:
7.129
TOGETHER WE REACH THE GOAL
IMPACT FACTOR:
7.429
ISI
INTERNATIONAL JOURNAL
FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY
Call : 08813907089 (24*7 Support on Whatsapp)


iJRASET
International Journal for Research in Applied Science & Engineering Technology
ISSN No. : 2321-9653
IJRASET is indexed with Crossref for DOI-DOI : 10.22214
Website : www.ijraset.com, E-mail : ijraset@gmail.com
ISRA Journal Impact Factor: 7.429
45.98
INDEX COPERNICUS
THOMSON REUTERS
TOGETHER WE REACH THE GOAL
SJIF 7.429
Certificate
It is here by certified that the paper ID : IJRASET59257, entitled
Lung Cancer Detection Using Machine Learning
by
C Ganesh
after review is found suitable and has been published in
Volume 12, Issue III, March 2024
in
International Journal for Research in Applied Science & Engineering Technology
(International Peer Reviewed and Refereed Journal)
Good luck for your future endeavors
Editor in Chief, iJRASET


iJRASET
International Journal for Research in Applied Science & Engineering Technology
ISSN No. : 2321-9653
IJRASET is indexed with Crossref for DOI-DOI : 10.22214
Website : www.ijraset.com, E-mail : ijraset@gmail.com
ISRA Journal Impact Factor: 7.429
45.98
INDEX COPERNICUS
THOMSON REUTERS
TOGETHER WE REACH THE GOAL
SJIF 7.429
Certificate
It is here by certified that the paper ID : IJRASET59257, entitled
Lung Cancer Detection Using Machine Learning
by
K Shiva Abhigna
after review is found suitable and has been published in
Volume 12, Issue III, March 2024
in
International Journal for Research in Applied Science & Engineering Technology
(International Peer Reviewed and Refereed Journal)
Good luck for your future endeavors
Editor in Chief, iJRASET


ISSN No. : 2321-9653
iJRASET
International Journal for Research in Applied Science & Engineering Technology
IJRASET is indexed with Crossref for DOI-DOI : 10.22214
Website : www.ijraset.com, E-mail : ijraset@gmail.com
ISRA Journal Impact Factor: 7.429
45.98
INDEX COPERNICUS
THOMSON REUTERS
Researcher ID: N-9681-2016
10.22214/IJRASET
TOGETHER WE REACH THE GOAL
SJIF 7.429
Certificate
It is here by certified that the paper ID : IJRASET59257, entitled
Lung Cancer Detection Using Machine Learning
by
P Tharun Sai
after review is found suitable and has been published in
Volume 12, Issue III, March 2024
in
International Journal for Research in Applied Science & Engineering Technology
(International Peer Reviewed and Refereed Journal)
Good luck for your future endeavors
Editor in Chief, iJRASET