International Journal of Leading Research Publication (IJLRP)
E-ISSN: 2582-8010 ● Website: www.ijlrp.com ● Email: editor@ijlrp.com

Lung Cancer Detection Using Machine Learning

Ms. Shubhangi Mahule[1], MD Tabrez[2], M.Prasanna[3], B. Praveen[4], B. Manisha[5]

[1, 2, 3, 4, 5]Computer Science and Engineering & ACE Engineering College, India

Abstract

Lung cancer is a dangerous disease that taking human life rapidly worldwide. The death of the people is increasing exponentially because of lung cancer. In order to reduce the disease and save a human's life, the automated system is needed. The purpose of the lung cancer detection system is able to detect and provide reliable information to doctors and clinicians from the medical image. To minimize this problem, many systems have been proposed by using different image processing techniques, machine learning, and deep learning techniques. A computed tomography (CT) imaging modality is an efficient technique for medical screening used for lung cancer detection and diagnosis. Physician and radiologist use the CT scan images to analyze, interpret and diagnose the lung cancer from lung tissues. However, in most cases, obtaining an accurate diagnosis result without using the extra medical tool known as a computer-Aid detection and Diagnosis (CAD) system is tedious work for many physicians. To obtain an accurate result from computer-aided diagnosis system lung segmentation methods are basic once. So in this project, we have used different lung segmentation and nodules segmentation methods. Our work has consisted of preprocessing, and lung segmentation by using thresholding, and also used the U-net model for detection of the candidate nodules of the patient's lung CT scan and classification methodology. We have used a convolutional neural network and designed a 3D CNN model that has a 0.77% accuracy performance.

Keywords: **Lung cancer, Cancer classification, Nodule detection, CNN, SVM, XG Boost Classifier**

I. INTRODUCTION

Lung cancer is one of the leading causes of cancer-related deaths worldwide, accounting for millions of fatalities annually. Early and accurate detection plays a crucial role in improving patient outcomes and enabling timely medical intervention. Traditional diagnostic methods, such as biopsies, X-rays, and computed tomography (CT) scans, are often time-consuming, expensive, and prone to human error. With the advancement of artificial intelligence (AI), machine learning (ML) has emerged as a revolutionary tool for automating and improving lung cancer detection. ML models can analyze vast amounts of medical data, identify patterns, and classify lung nodules with high accuracy, assisting radiologists and oncologists in early diagnosis. This paper explores the role of ML in lung cancer detection, focusing on various learning models, datasets, challenges, and future directions. The integration of ML in medical imaging and diagnostics is expected to enhance accuracy, reduce misdiagnosis, and ultimately improve patient survival rates.

II. LITERATURE SURVEY

[1]. The study titled "An Extensive Review on Lung Cancer Detection Using Machine Learning Techniques: A Systematic Study(2020)" Investigates the accuracy levels of various machine learning algorithms for lung cancer nodule detection. Conducted a systematic literature review to evaluate different models employed by researchers. Identified limitations and drawbacks of existing methods. Found that some classifiers have low accuracy, while others have higher accuracy but are not perfect. Discovered that improper handling of DICOM images contributed to low accuracy in some cases. Ensemble classifiers demonstrated superior performance compared to other machine learning algorithms.[1]

[2]. The study titled "A Review of Most Recent Lung Cancer Detection Techniques Using Machine Learning (2021)" advancements in lung cancer detection techniques using machine learning. Key points include: Lung cancer is a leading cause of death, and early diagnosis can significantly improve survival rates. CT scans are found to be more effective than X-rays for lung cancer detection. Various classifiers like Support Vector Machine (SVM), Random Forest, Decision Tree, Neural Networks, and Convolutional Neural Networks (CNN) are analyzed. Marker-controlled watershed segmentation is identified as the most effective segmentation method. Deep learning techniques, particularly CNN-based models, achieve the highest accuracy (up to 99%).

[3]. The study titled "An Extensive Review on Lung Cancer Diagnosis Using Machine Learning Techniques on Radiological Data: State-of-the-art and Perspectives(2023)" Lung cancer is one of the leading causes of cancer-related deaths globally, with early detection being vital for increasing survival rates. The study emphasizes the role of automated diagnosis systems, using machine learning (ML) and deep learning (DL) techniques, to analyze radiological data (CT, MRI, X-rays). A decade-long review of research from databases such as Deep learning, particularly Convolutional Neural Networks (CNNs), has been revolutionary in medical imaging. Support Vector Machines (SVM) and K-nearest neighbors (KNN) are also effective in classifying lung cancer. Challenges remain, such as low accuracy with certain algorithms and limited datasets.

[4]. The study titled "Enhancing Lung Cancer Detection Through Hybrid Features and Machine Learning Hyperparameters Optimization Techniques(2024)" The study focuses on improving lung cancer detection using hybrid feature extraction methods and optimized machine learning techniques. Key contributions include. Integration of features such as Gray-Level Co-occurrence Matrix (GLCM), Haralick texture, and autoencoder-based features. Implementation of supervised machine learning models like Support Vector Machine (SVM) with Radial Basis Function (RBF), Gaussian, and Polynomial kernels. Achieving high detection accuracy by fine-tuning hyperparameters through grid search. Application of 10-fold crossvalidation and data augmentation to enhance model reliability. Key results demonstrated 100% accuracy using hybrid features (e.g., GLCM + Autoencoder or Haralick + Autoencoder) with SVM RBF and Gaussian models

[5]. Emre Dandil has been proposed a computer-aid pipeline for automatic lung cancer classification on Computed Tomography (CT) scan. The system used a private dataset that consists of 47 CT scans from 47 different patients. This proposed pipeline has composed on four stages: (1) Image preprocessing stage, in this stage, CT scan images are enhanced, and lung volume are extracted from the image, (2) Nodule detection stage. (3) Feature computation stage that used to extract features from lung image, and Principal Component Analysis (PCA) is used for feature reduction. The final stage is classification,

in this stage the system, in this stage, the system has been used Probabilistic Neural Network (PNN) that used for benign and malign nodules.

III. EXISTING SYSTEM

The existing systems for lung cancer detection primarily rely on manual diagnostic methods such as physical examination, X-rays, CT scans, and biopsies, which require expert interpretation by radiologists and oncologists. These methods are time-consuming, prone to human error, and often fail to detect cancer in its early stages when symptoms are minimal or absent. The lack of automation and real-time analysis in current systems limits their efficiency and delays critical medical intervention, thereby impacting patient outcomes.

In existing systems, the diagnostic workflow for lung cancer typically begins with imaging techniques such as chest X-rays or CT scans, followed by further clinical evaluation and invasive procedures like biopsies for confirmation. These methods are dependent on the availability and expertise of radiologists and pathologists, which may not be consistently accessible in all healthcare settings, especially in rural or under-resourced regions. Additionally, the interpretation of medical images is subjective and may vary between practitioners, leading to inconsistent results and delayed or incorrect diagnoses.

Another limitation of the current systems is their inability to handle large volumes of patient data efficiently. As the number of diagnostic cases increases, manual analysis becomes a bottleneck, potentially leading to increased workload, fatigue, and diagnostic oversight. Moreover, existing tools often lack integration with modern computational technologies that can assist in predictive analytics or pattern recognition. This absence of intelligent support tools not only affects the speed and accuracy of diagnosis but also limits opportunities for early intervention, which is crucial in improving survival rates for lung cancer patients.

IV. PROPOSED SYSTEM

The proposed system architecture for lung cancer detection integrates multiple machine learning (ML) models with an image-based deep learning approach to deliver accurate predictions through a web interface. The process begins with a dataset that undergoes preprocessing, including normalization and feature extraction, to prepare the data for modeling. Various ML algorithms such as Logistic Regression, Support Vector Machine (SVM), and K-Nearest Neighbor (KNN) are then applied to the processed data. These models are trained and evaluated to determine the most effective one, which is selected as the best model for deployment. In parallel, a Convolutional Neural Network (CNN) is employed to handle image-based data, particularly CT scan images, allowing for more precise detection of lung nodules and cancer indicators.

Once the best-performing model is identified, it is integrated into a user-friendly application using the Streamlit framework. This enables real-time predictions through a web interface, making the system accessible and practical for clinical environments. The Streamlit interface accepts CT scan images as input and displays the prediction results to the user, thereby reducing the need for manual interpretation and minimizing human error. The combination of traditional ML models for structured data and CNNs

for image analysis, along with seamless web deployment, enhances the system's overall effectiveness in early and reliable lung cancer detection.

This hybrid approach not only leverages the strengths of both machine learning and deep learning but also ensures scalability and usability through web integration. By streamlining the diagnostic process and offering accurate, automated predictions, the system has the potential to support radiologists in making faster and more informed decisions, ultimately contributing to improved patient care and early intervention in lung cancer cases.

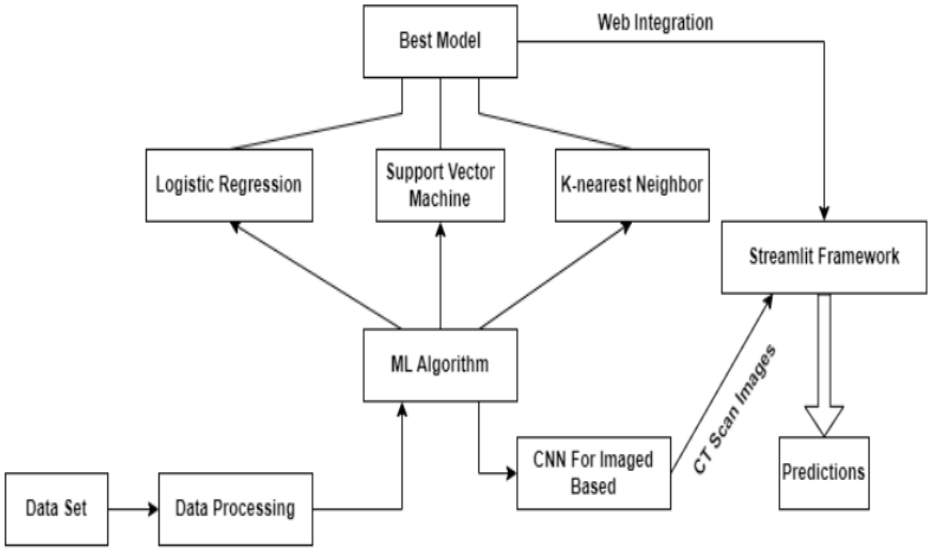

Fig. 4.1 System Architecture

V. METHODLOGIES

5.1 Convolutional Neural Networks (CNNs) are widely used in medical image analysis, particularly for detecting lung cancer in CT scans and X-rays. They consist of multiple layers, including convolutional layers that extract features such as edges, textures, and tumor shapes. Pooling layers help reduce dimensionality, making computations more efficient while retaining important features. Fully connected layers then classify images into cancerous or non-cancerous categories. CNNs excel in capturing spatial patterns, allowing them to differentiate between benign and malignant tumors with high accuracy. Transfer learning techniques, where pre-trained CNN models like VGG16 or ResNet are fine-tuned on medical datasets, further enhance detection performance. A major advantage of CNNs is their ability to learn from raw image data without manual feature extraction. However, they require large amounts of labeled data and high computational power.

5.2 The K-Nearest Neighbors (KNN) algorithm is a simple yet effective machine learning approach used in lung cancer detection. It is a non-parametric, instance-based learning algorithm that classifies a new data point based on the majority class of its nearest neighbors. The value of 'K' determines how many neighbors influence the classification, with larger values leading to smoother decision boundaries. KNN is particularly useful in diagnosing lung cancer by comparing new patient data, such as genetic markers or symptoms, to previously labeled cases. It works well for small datasets where real-time

predictions are not required. However, KNN becomes computationally expensive for large datasets since it requires storing and searching through the entire dataset during prediction. Additionally, the choice of distance metric, such as Euclidean or Manhattan distance, plays a crucial role in determining accuracy.

5.3 XGBoost is a powerful machine learning algorithm known for its efficiency and high predictive accuracy in lung cancer detection. It is a gradient boosting algorithm that builds multiple weak decision trees sequentially, improving performance with each iteration by correcting previous errors. One of its key advantages is its ability to handle structured medical data, such as patient history, lab results, and tumor biomarkers. XGBoost is optimized for speed and performance, using parallel computing and regularization techniques to prevent overfitting. It efficiently handles missing values and categorical variables, making it ideal for lung cancer diagnosis where datasets may contain incomplete records. Additionally, it provides feature importance scores, helping researchers understand which factors contribute most to cancer risk. Despite its advantages, XGBoost requires careful hyperparameter tuning for optimal results.

5.4 Support Vector Machine (SVM) is a powerful supervised learning algorithm widely used for lung cancer detection due to its ability to handle high-dimensional data. SVM works by finding the optimal hyperplane that best separates different classes, such as cancerous and non-cancerous cases, based on input features like tumor size, shape, and texture. It uses a kernel trick to transform non-linearly separable data into a higher-dimensional space, making it effective for complex medical datasets. Common kernels include linear, polynomial, and radial basis function (RBF), each suited for different types of data distributions. SVM is particularly useful for small to medium-sized datasets where clear class boundaries exist. However, its performance can be affected by the choice of hyperparameters, requiring careful tuning for optimal results. It is computationally intensive for large datasets but remains a reliable choice for medical diagnosis due to its robustness against overfitting.

5.5 Random Forest (RF) is a widely used ensemble learning algorithm that improves lung cancer detection by combining multiple decision trees. Each tree in the forest is trained on a random subset of the data, and the final prediction is made by aggregating the outputs of all trees, reducing the risk of overfitting. This approach enhances model stability and accuracy, making RF effective for medical datasets that contain diverse patient information, including imaging features, genetic markers, and clinical history. RF handles missing values well and provides feature importance scores, helping identify key risk factors for lung cancer. It is particularly useful in diagnosing lung cancer from structured data, such as patient records, rather than imagebased analysis. Although RF is computationally efficient compared to deep learning models, it may struggle with interpretability when dealing with complex decision boundaries. Despite this, its high accuracy, robustness, and ease of implementation make it a preferred choice in medical diagnostics, assisting doctors in making informed decisions.

VI. RESULTS

6.1 HEAT MAP

A **heat map** in the context of lung cancer detection using machine learning is a visual representation that highlights areas of a lung scan (such as a CT image) where a model identifies features indicative of cancer. It assigns different colors—typically warmer colors like red or orange for high attention and cooler colors like blue for low attention—to show the regions that most influence the model's prediction. Heat maps are commonly generated using techniques like **Grad-CAM (Gradient-weighted Class Activation Mapping)** in convolutional neural networks (CNNs) and are essential for interpretability, helping radiologists understand which areas of the lung the model focused on when making a decision. This not only builds trust in AI systems but also assists in verifying and validating the detection results.

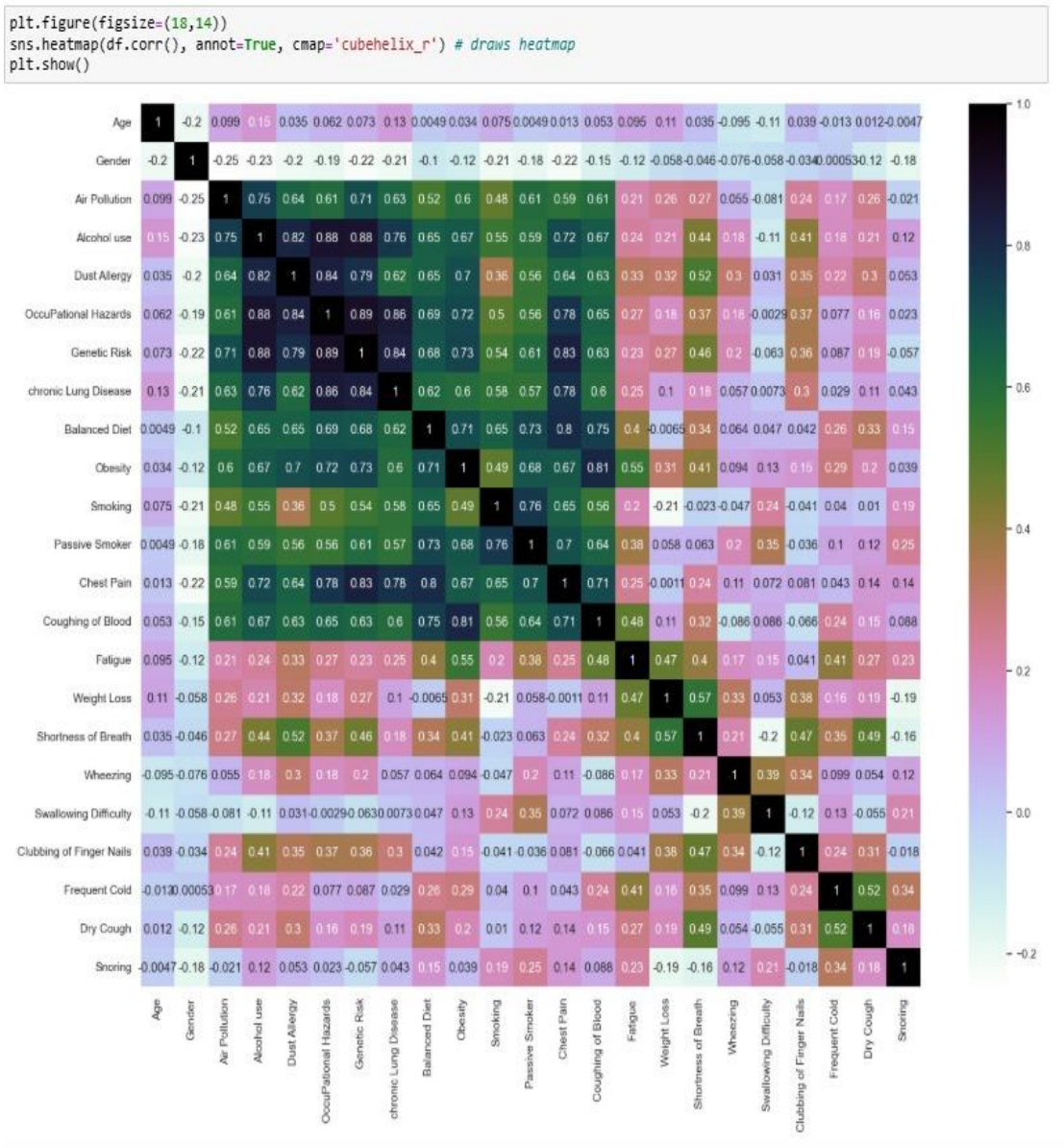

Fig. 6.1 Heatmap

6.2 USER INTERFACE

The front end of the Lung Cancer Detection System has been developed using **Streamlit**, a powerful and easy-to-use Python framework for building interactive web applications. The interface features a clean and intuitive layout with a sidebar navigation menu that guides users through different sections such as *Introduction*, *About the Dataset*, *Lung Cancer Prediction*, and *CNN-Based Disease Prediction*. The main display area is designed to showcase results and visual insights, such as annotated lung scans. This user-friendly design enables both medical professionals and researchers to interact with the prediction models in real-time, simplifying the diagnostic workflow and enhancing decision-making in clinical settings.

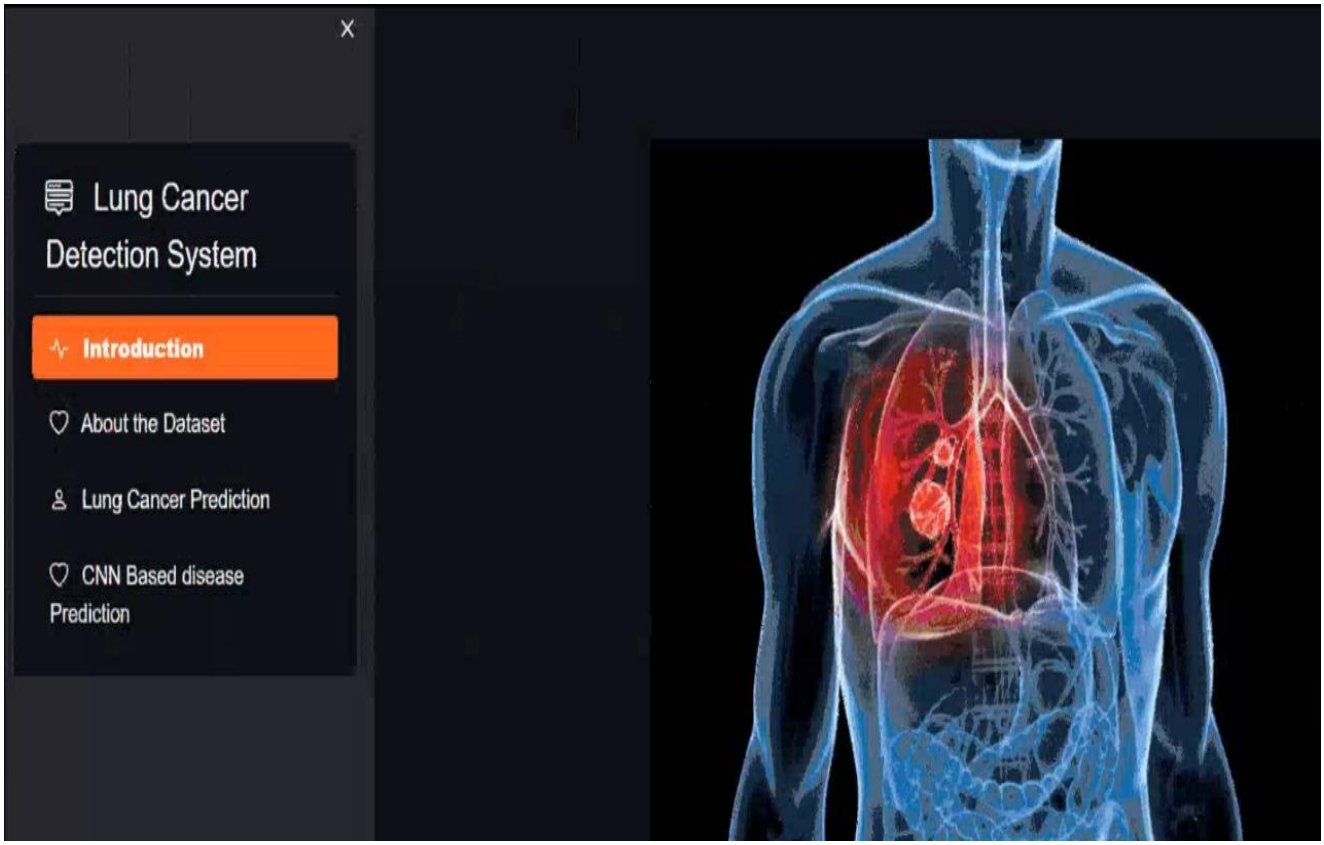

Fig. 6.2 User Interface

6.3 PREDICTION OUTPUT

The uploaded image displays a computer screen showing a user interface for uploading CT scan images. A CT scan of the lungs is visible at the bottom of the screen, and above it, a green banner proudly states, "I am 99.99 percent confirmed that this is a Normal Case." This suggests an automated system is analyzing the CT scan and providing a high-confidence assessment that the lungs appear healthy. The interface also shows file upload options, indicating this could be part of a medical imaging analysis application or a diagnostic tool.

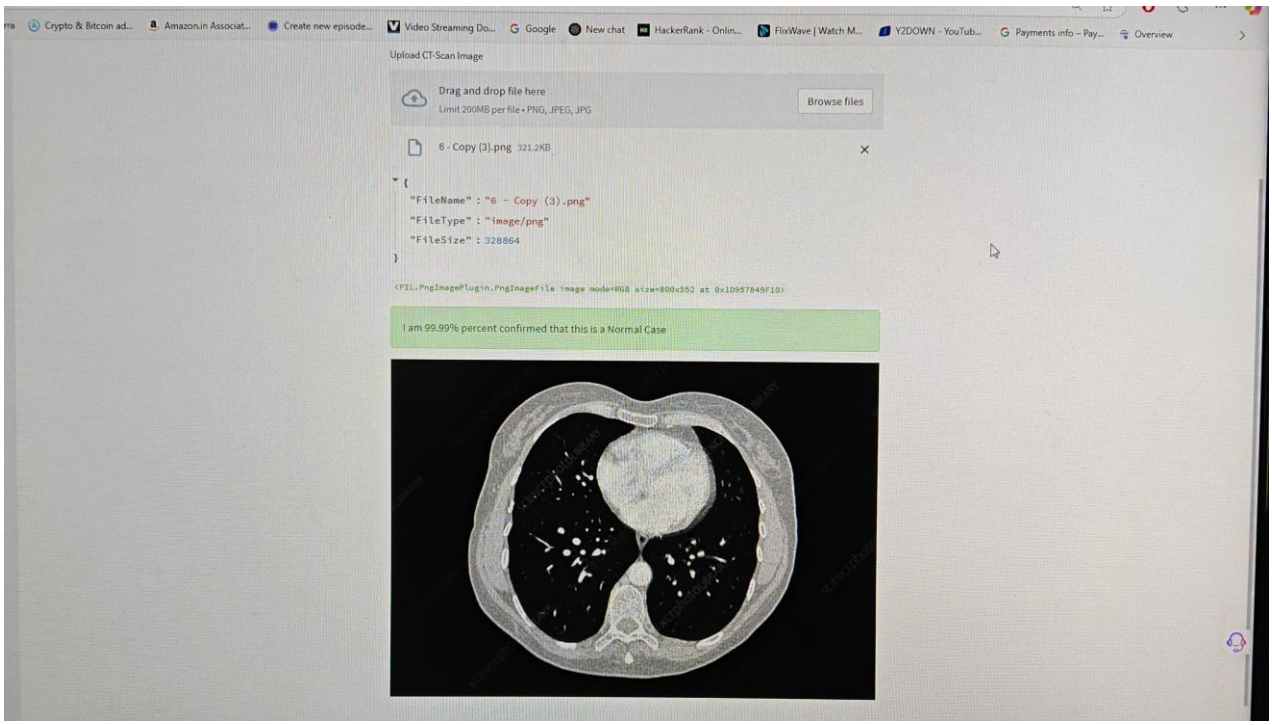

Fig. 6.3 Prediction Output

VII.CONCLUSION

In conclusion, lung cancer detection using machine learning marks a transformative step in medical diagnostics, enabling faster, more accurate, and data-driven identification of cancerous conditions. By employing a combination of algorithms such as Support Vector Machine (SVM), Random Forest (RF), Naive Bayes, and k-Nearest Neighbors (KNN), the system effectively analyzes complex medical data—including CT scans and X-rays—to identify early signs of lung cancer with high precision. These techniques complement each other to enhance robustness and reliability, while the system's ability to adapt and improve with new data ensures long-term effectiveness. Moreover, automation reduces human error, streamlines the diagnostic workflow, and facilitates timely medical interventions. Ultimately, the integration of machine learning not only supports healthcare professionals in making informed decisions but also holds great promise for improving patient outcomes and combating lung cancer more effectively.

REFERENCES

[1]. LUNG CANCER DETECTION DATASET: A MEDICAL IMAGING-BASED APPROACH - ADITYA MAHEMAKAR, MOSTAPHAGANEM , SUSENA VENKATESH

[2]. Lung Cancer Diagnosis and Early Detection: AI-Based Approaches - Zainab Gandi , Priyatam Gurram, Birendra Admai

[3]. AI-Based Lung Cancer Detection System Using Deep Learning – Mohammad ,Qusai Abuein , Romesa Al-Quren

[4]. A Big Data-Based Predictive Analytics System for Lung Cancer Detection – Ravi B Parik, Andrew Gdowski,Justin K

[5]. Facing Challenges in AI-Powered Cancer Diagnosis – Willer Mark, Gellerstedt

[6]. How Effective Are AI-Assisted Cancer Diagnosis Tools? Wenya Linda B, Ahemd Hosny, Mathew B
[7]. Effect of Using AI for Lung Cancer Diagnosis: A Comparative Analysis - Michaela Celliva, Laura Maria Cacippa
[8]. Using Augmented Reality to Stimulate Students and Diffuse Escape Game Activities to Larger Audiences - Amir H. Sadeghi, Quinten Mank, Alphers
[9]. Mapping Research in AI-Based Cancer Detection and Precision Medicine - Bhavnet Bhinder, Coryander Gilvary, Neel S Madhukar
[10]. An Extensive Review on Lung Cancer Detection Using Machine Learning Techniques: A Systematic Study - Eali ,Stephen Neal, Debnath Bhattacharyya2
[11]. An Extensive Review on Lung Cancer Diagnosis Using Machine Learning Techniques on Radiological Data: State-of-the-art and Perspectives - Syed Naseer Ahmad Shah, Rafat Parveen
[12]. Enhancing Lung Cancer Detection Through Hybrid Features and Machine Learning Hyperparameters Optimization Techniques - Liangyu Li, Jing Yang, Lip Yee Por, Mohammad Shahbaz Khan
[13]. A Review of Most Recent Lung Cancer Detection Techniques Using Machine Learning - Dakhaz Mustafa Abdullah & Nawzat Sadiq Ahmed
[14]. Machine Learning-Based Lung Cancer Detection Using Multiview Image Registration and Fusion - Imran Nazir,Ihsan ul Haq, Mostafa Dahshan ,Muhammad Mohsin Jadoon
[15]. Deep learning for lungs cancer detection: a review - Rabia Javed ,Tahir Abbas,Ali Haider Khan, Ali Daud,Amal Bukhari