Cancer Sentinel Using Machine Learning to Improve Lung Tumour Identification

[1]ALEKHYA.T, [2]ARUMALLA SAI KIRAN REDDY, [3]PIDUGURALLA VENKATA SAI, [4]PAPASANI AKASH, [5]DASI SANDEEP

[1]Assistant Professor, Department of CSE-AI&ML, Kallam Haranadhareddy Institute of Technology, Guntur, AP, India.

[2,3,4,5] B.Tech Students, Department of CSE-AI&ML, Kallam Haranadhareddy Institute of Technology, Guntur, AP, India.

Abstract

In the domain of lung tumor detection, several existing systems have employed various techniques such as Support Vector Machines (SVM), Random Forest, and image fusion methods. While these methods have shown promise, they are not without their limitations, which hinder their effectiveness in addressing the complexities of lung tumor detection.

Support Vector Machines (SVM) have been widely used in medical image analysis due to their ability to handle high-dimensional data and their effectiveness in binary classification tasks. However, SVMs may struggle with multi-class classification problems, such as distinguishing between normal lung tissue, benign tumors, and malignant tumors, which are prevalent in lung tumor detection. Additionally, SVMs are sensitive to the choice of kernel function and require careful parameter tuning, making them less suitable for complex datasets with heterogeneous features.

Random Forest is another popular machine learning algorithm used for classification tasks. It works by constructing multiple decision trees during training and outputs the mode of the classes for classification problems. While Random Forests are robust to overfitting and can handle nonlinear relationships between features and labels, they may not perform optimally with highly imbalanced datasets, such as those encountered in lung tumor detection. Moreover, Random Forests may struggle to capture the spatial dependencies and intricate patterns present in medical images, which are crucial for accurate tumor detection.

The proposed system offers several benefits over existing methods. Firstly, by utilizing CNNs, it can effectively capture intricate patterns and spatial dependencies in lung images, leading to more accurate and reliable tumor detection. Secondly, the integration of SMOTE and class weight adjustments helps mitigate the challenges associated with imbalanced datasets, resulting in a more balanced and robust classification model. Thirdly, the proposed system has the potential to outperform traditional machine learning algorithms such as SVM and Random Forest, which may struggle with complex image data and imbalanced class distributions.

Overall, the proposed system presents a promising approach for lung tumor detection, addressing the limitations of existing methods and offering improved performance, accuracy, and reliability in the diagnosis of lung cancer.

I. INTRODUCTION

Lung cancer is a prevalent and potentially life-threatening disease that originates in the cells of the lungs. It is a type of cancer characterized by uncontrolled cell growth in the tissues of the lungs, leading to the formation of tumors. The primary risk factor for lung cancer is tobacco smoke, which contains numerous carcinogens that can damage lung cells over time. However, non-smokers can also develop lung cancer due to exposure to secondhand smoke, environmental pollutants, and genetic factors. Additionally, individuals with a family history

of lung cancer or those with a personal history of lung diseases, such as chronic obstructive pulmonary disease (COPD), may be at an increased risk Preventing lung cancer involves adopting lifestyle choices and practices that reduce exposure to known risk factors such as Smoking Cessation, Avoiding Second hand smoke, Occupational safety, and Healthy lifestyle choices.

These tumors can be categorized into three main types: normal, benign, and malignant.

Normal lung tissue: Normal lung tissue refers to the healthy cells and structures that make up the lungs. Normal lung tissue does not exhibit any abnormal growth or formation of tumors

Fig 1: Normal Lung Tissue

Benign lung tumors: Benign lung tumors are non-cancerous growths that develop within the lung tissue. These tumors typically grow slowly and do not spread to other parts of the body. While benign tumors may cause symptoms such as coughing, wheezing, or chest pain, they are usually not life-threatening and can often be successfully treated through surgical removal or other minimally invasive procedures.

Fig 2 : Benign Lung Tumor

Malignant lung tumors: Malignant lung tumors, commonly referred to as lung cancer, are cancerous growth that originate in the lung tissue. These tumors can be further classified into two main types: non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC)

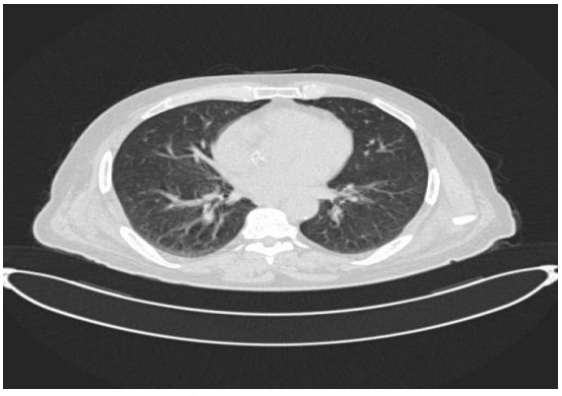


Fig 3 : malignant Lung Tumor

Proposed System

The proposed system for lung tumor detection utilizes Convolutional Neural Networks (CNNs), augmented with Synthetic Minority Over- sampling Technique (SMOTE) and class weight adjustments to address theshortcomings of existing methods such as Support Vector Machines (SVM), Random Forest, and image fusion techniques. CNNs are particularly well-suited for image classification tasks due to their ability to automatically learn hierarchical representations of features from raw pixel data. Unlike traditional machine learning algorithms, CNNs can capture complex spatial patterns and dependencies present in medical images, making them more effective for detectingsubtle abnormalities like lung tumors. By leveraging deep learning techniques, the proposed system aims to improve the accuracy and robustness of lung tumor detection compared to previous methods.

To mitigate the challenges posed by imbalanced datasets, SMOTE is employed to generate synthetic samples of minority classes (benign and malignant tumors), thereby balancing the distribution of tumor classes in the training data. This helps prevent bias towards the majority class (normal lung tissue) and ensures that the CNN model learns to accurately distinguish between different tumor types.

Additionally, class weight adjustments are applied during training to assign higher weights to minority classes, ensuring that the model pays more attention to correctly classifying rare tumor types. This helps addressthe imbalance issue encountered in medical imaging datasets and improves the overall performance of the detection system.

The proposed system offers several benefits over existing methods. Firstly, by utilizing CNNs, it can effectively capture intricate patterns and spatial dependencies in lung images, leading to more accurate and reliable tumor detection. Secondly, the integration of SMOTE and class weight adjustments helps mitigate the challenges associated with imbalanced datasets, resulting in a more balanced and robust classification model. Thirdly, the proposed system has the potential to outperform traditional machine learning algorithms such as SVM and Random Forest, which may struggle with complex image data and imbalanced class distributions.

Unique features of the System

The proposed system for lung tumor detection introduces several distinctive features that collectively enhance its efficacy and applicability in medical imaging analysis. At its core, the system leverages Convolutional Neural Network (CNN) architecture specifically tailored for medical image interpretation. Unlike conventional methods,CNNs autonomously learn intricate patterns and spatial dependencies inherent in lung images, enabling precise tumor detection with unprecedentedaccuracy.

A notable innovation of the proposed system lies in its integration of Synthetic Minority Over-sampling Technique (SMOTE) for data augmentation. This approach effectively addresses the inherent challenge of imbalanced datasets encountered in medical imaging, particularly in lung tumor detection. By synthetically generating samples of minority classes—such as benign and malignant tumors—SMOTE rebalances the dataset, ensuring that the model is equally trained on all tumor types. This significantly improves the model's ability to generalize and accurately identify rare tumor instances, which might otherwise be overlooked or misclassified.

Furthermore, the system implements class weight adjustments during training, prioritizing the correct classification of minority classes by assigning higher weights. This strategic weighting mechanism mitigates bias towards the majority class (normal lung tissue) and enhances the overall robustness and reliability of the detection model. Additional settings and imaging modalities. Designed to seamlessly integrate into existing diagnostic workflows, the system can be tailored to accommodate specific clinical requirements and

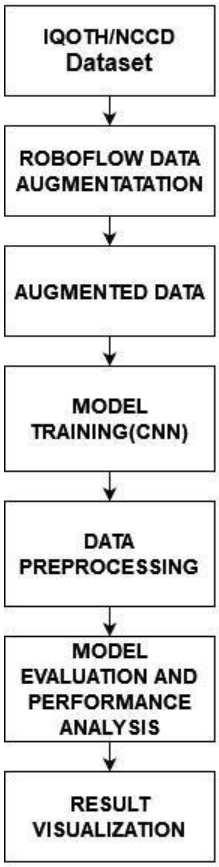
preferences. Itsversatility facilitates widespread deployment in various clinicalenvironments, empowering healthcare professionals with advanced tools for early An additional advantage of the proposed system is its scalability and adaptability across diverse healthcare diagnosis, treatment planning, and patient monitoring. Moreover, the proposed system embodies a commitment to continual learning and improvement. By incorporating user feedback and updated data, it evolves iteratively, adapting to emerging challenges and advancements in medical imaging technology. This dynamic approach ensures that the system remains at the forefront of lung tumor detection, consistently delivering optimal performance andreliability in clinical practice.

II. LITERATURE REVIEW

In 2019, Moradi compared different techniques to differentiate lung cancer nodules from non-nodules. To reduce/eliminate the false positive predictions they have come up with 3D Convolutional Neural Network Technique. Nodules exist in different sizes and using just one CNN can result in false detections. So they divided the nodules into four groups according to their size. And they have used four different sizes of 3D CNN. They combined all those 4 classifiers to get better results. EachCNN consists of a nubmber of 3D CNN which are all varying sizes. All 4 classifiers were combined in order to produce results which were better. A combination of Max pooling layer and convolutional layer were used to produce each CNN. The activation function used here is ReLU. Softmax layer accompanied by a fully connected layer is used to produce the output finally. Nodules size varies from 3mm to 3cm so by using just one layer, the prediction could be wrong for either very small nodules or very large values. So they fused all the 4 CNNs and sent their output values (predicted values) to a final classifier.In 2018, Bohdan Chapliuk et al. [4] applied neural networks C3D and 3D DenseNet to detect lung cancer usingCT images. These Neural networks were applied to whole lung 3D images and two-stage approaches (for segmentation and classification, two different neural networks are trained.) and further compared. Data Science Bowl 2017 dataset containing CT scans of more than 1000 patients was used. For pre-processing all the CT images were converted into HouseholdUnits (HU is a unit describing x-ray intensity) by resampling. HU ranges are specific to tumors (-500) so, in the second step, a range for lung tissue that filters out all bones from the image was filtered out by all patient images.

The size of the 3D patient image was reduced to 120x120x120.In 2020, QINGHAI ZHANG et al. proposed a method for designing of Lungnodule detection system which is automatic. The dataset used for the proposed method is LIDC-IRDI public dataset. The proposed method usedfor this study is Multi-Scene Deep Learning Framework which contains several steps. CT images are given as input and the probability distributionof distinct gray levels is obtained by threshold segmentation that is Histogram. Correcting the smooth lung outlines is the main aim for the lung parenchyma segmentation process. The replacement of the vein system in the lung helps to identify the nodule structure. Vessel filters are used for removing the vessels which reduce the number of false positive. The design of CNN contains a pooling layer, a convolutional layer, and a fully integrated layer. Segmentation and classification identify Class 1 and Class2 that are two class of image data and discrete images which are separated from the lung images respectively. Finally, zero-centering was achieved by subtracting the mean value of the images from the training dataset. A U-Net was trained using the LUNA16 dataset instead of entering the segmented images directly into the classifier, to detect the exact position of nodules. Accuracy, false-positive rate, Mis-Classificationrate and false-negative rate were found to be 86.6%, 11.9%, 13.4% and 14.7% respectively.

III. DESIGN OF THE SYSTEM



IQOTH/NCCD Dataset:
The source dataset containing medical imaging data for lung tumor detection.
Data Preprocessing Pipeline:
A series of preprocessing steps to clean, normalize, and prepare the data for model training. This includes tasks such as image resizing, normalization, and annotation parsing.
Roboflow Data Augmentation:
Utilizes Roboflow to apply augmentation techniques such as rotation,flipping, and color jittering to enhance dataset variability and improve model generalization.
Augmented Data:
The output of the data preprocessing and augmentation pipeline, containing augmented and labeled images ready for model training.
Model Training (CNN):
The Convolutional Neural Network (CNN) architecture trained on theaugmented dataset to learn features and patterns indicative of lung tumors. **Model Evaluation & Performance Analysis:**
Assessing the trained model's performance using evaluation metrics such as accuracy, precision, recall, and F1-score. Performance analysis helps identify areas for improvement and ensures the model meets desired criteria.
Result Visualization:
Visualization of model outputs, including predicted tumor locations, confidence scores,and diagnostic findings, to aid clinicians in interpreting and validating the model's predictions.

Architectural Design

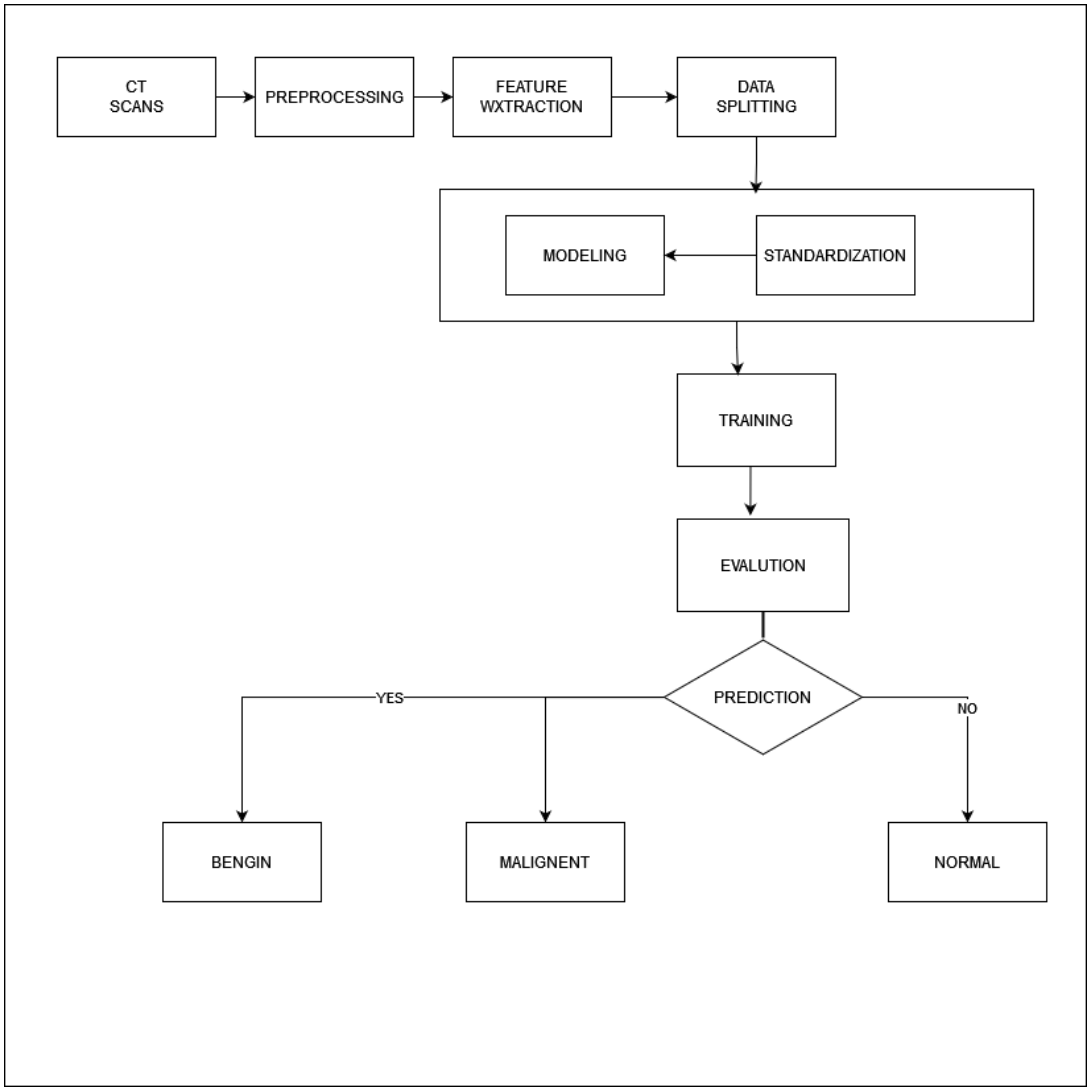


Fig : Architecture of Identification Of Lung Tumor Through ML

The architecture begins with the collection of medical imaging data from the IQOTH dataset, which contains chest X-rays or CT scans with annotations indicating tumor locations.This data serves as the foundation for model training and evaluation. The collected data undergoes augmentation using techniques such as rotation, flipping, and scaling to increase dataset variability and improve model generalization. Roboflow is utilized for automated augmentation, ensuring efficient generation of augmented training data**.**

The augmented data is then used to train a Convolutional Neural Network (CNN) model, selected based on its suitability for medical image analysis. The CNN architecture is customized and optimized to detect lung tumors effective The training phase involves feeding the augmented dataset into the CNN model to learn features and patterns indicative of lung tumors. Techniques such as transfer learning and fine-tuning may be employed to leverage pre-trained models and accelerate training.

Following training, the model is evaluated using a separate test dataset to assess its performance. Evaluation metrics such as accuracy, precision, recall, and F1-score are computed to gauge the model's effectiveness in tumor detection.

The testing results are analyzed to identify areas of strength and improvement in the model's performance. Visualization tools may be utilized to interpret model predictions and highlight

regions of interest within the medical images.
Once the model has been trained and evaluated, it is deployed within a Stream lit application, providing clinicians with a user-friendly interface for uploading medical images and receiving tumor predictions. This deployment enables real-time interaction and decision-making based onthe model's outputs.

Algorithms Design

Algorithms are very important in computer Science. The best-chosen algorithm makes sure the computer will do the given task in the best possible manner. In cases where efficiency matters a proper algorithm is
really vital to be used. An algorithm is important in optimizing a computerprogram according to the available resources.
Step 1: Data Collection
Step 2: Data Augmentation
Step 3: Data Preprocessing
Step 4: Model Training
Step 5: Model Evaluation
Step 6: Performance Analysis
Step 7: Deployment

Step 1:

Data Acquisition:

Collect chest X-ray or CT scan images along with annotations indicatingtumor locations. Use a dataset such as IQOTH.

Step 2:

Augmentation:

Apply augmentation techniques such as rotation, flipping, and scaling to increase dataset variability and improve model generalization. Use tools like Roboflow for automated augmentation.

Step 3:

Preprocessing:

Load the chest X-ray or CT scan images along with their correspondingannotations.
Perform any necessary preprocessing steps such as resizing, normalization,and augmentation to prepare the data for training.

Step 4:

Model Selection:

Choose a suitable machine learning model for medical image analysis,such as Convolutional Neural Networks (CNNs).

Customization:

Customize the CNN architecture to detect lung tumors effectively. This may involve adjusting the number of layers, filter sizes, and activation functions.

Training:

Feed the augmented dataset into the CNN model to learn features and patterns indicative of lung tumors. Employ techniques like transfer learning and fine-tuning if necessary to leverage pre-trained models and accelerate training.

Step 5:

Test Dataset:

Separate a portion of the dataset for testing purposes.

Evaluation Metrics:

Evaluate the trained model using evaluation metrics such as accuracy, precision, recall, and F1-score to assess its performance in tumor detection.

Step 6:
Analysis:
Analyze the testing results to identify areas of strength and improvement in the model's performance. Utilize visualization tools to interpret model predictions and highlight regions of interest within the medical images.
Step 7:
Application Development:
Develop a user-friendly application (e.g., using Streamlit) for clinicians toupload medical images and receive tumor predictions.
Deployment:
Deploy the trained model within the application, enabling real-timeinteraction and decision-making based on the model's outputs.

IV. RESULTS

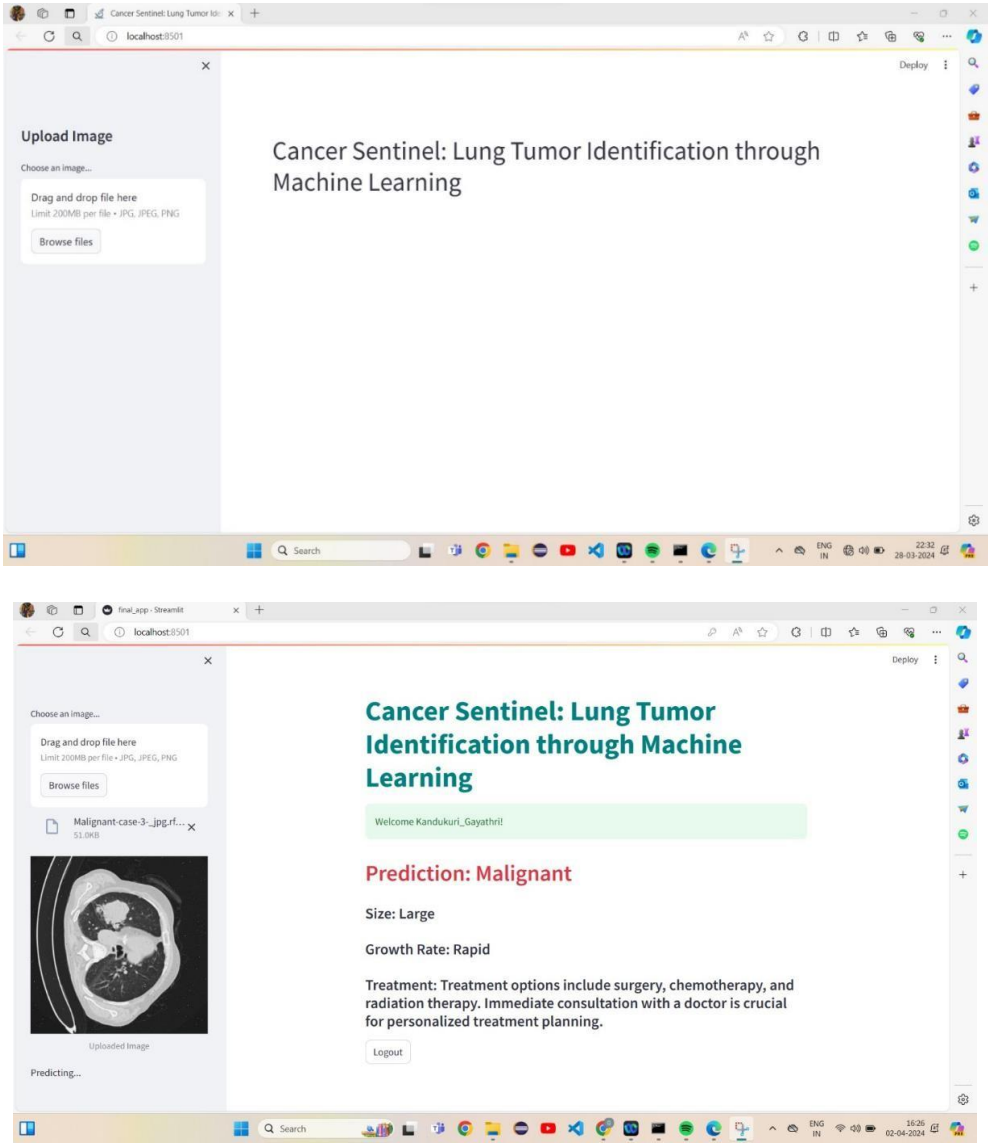


Fig 5: Webpage when we uploaded Malignant CT Scan

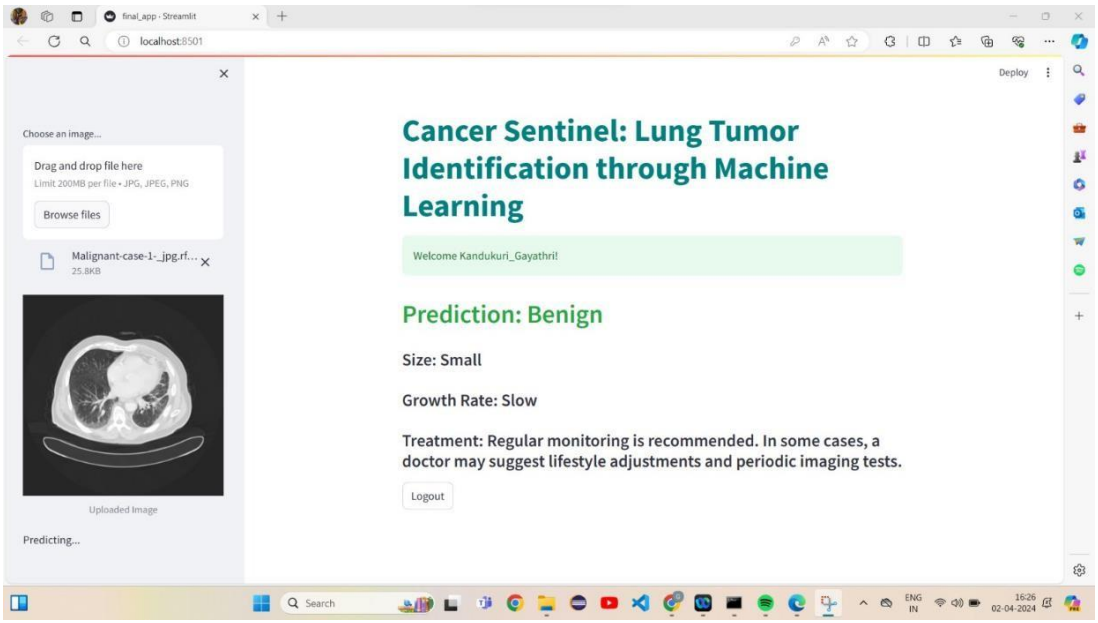


Fig 6 : Webpage when we uploaded benign CT Scan

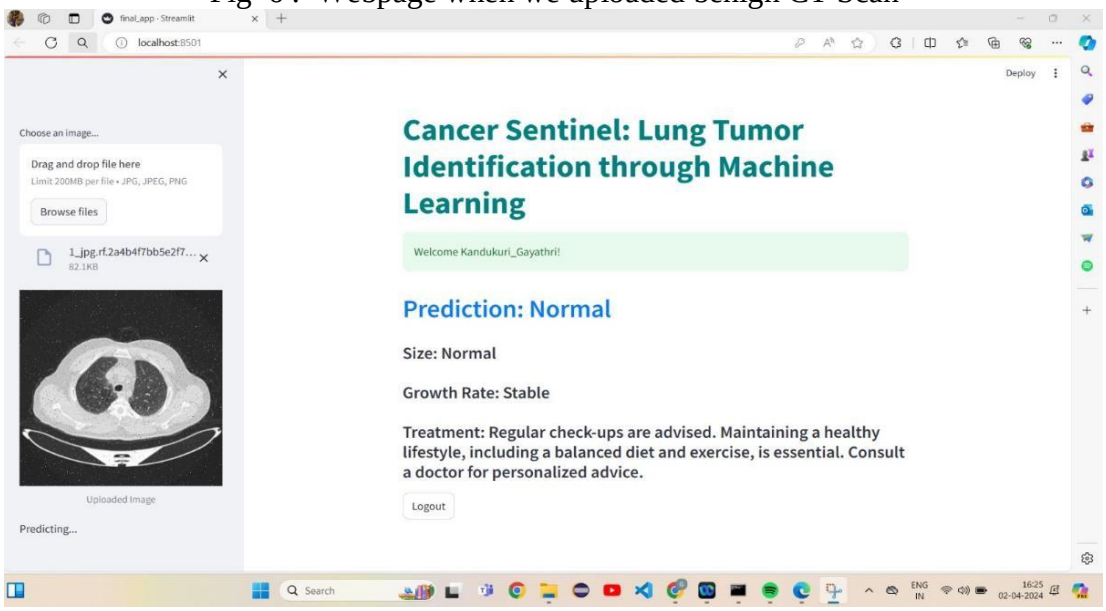


Fig 7: Webpage when we uploaded Normal CT Scan

V. CONCLUSION AND FUTURE ENHANCEMENTS

CONCLUSION

The lung tumor detection project represents a significant advancement in leveraging machine learning and deep learning techniques to assist in the diagnosis and management of lung tumors. By integrating advanced technologies with medical imaging data, this project aims to improve the accuracy, efficiency, and accessibility of lung tumor detection processes.

Through the development and deployment of a convolutional neural network (CNN) model trained on a diverse dataset of lung tumor images, the project provides a valuable tool for healthcare professionals to aid in the early detection and characterization of lung tumors. The implementation of data augmentation, class weighting, and model optimization techniques enhances the robustness and reliability of the system, enabling accurate predictions across different tumor types and imaging conditions.

The Streamlit-based user interface offers a user-friendly platform for uploading medical

images, receiving real-time tumor predictions, and accessing diagnostic recommendations. This interface facilitates seamless interaction with the system, empowering healthcare professionals to make informed decisions and provide timely interventions for patients.
Furthermore, thorough testing methodologies ensure the functionality, accuracy, performance, and security of the system, thereby instilling confidence in its reliability and effectiveness. By conducting unittesting, integration testing, functional testing, performance testing, and user acceptance testing, developers ensure that the system meets the highest standards of quality and usability. Overall, the lung tumor detection project holds immense potential to revolutionize the field of oncology by streamlining the diagnostic process, improving patient outcomes, and ultimately saving lives. Through continued research, development, and collaboration with medical experts, this project serves as a testament to the transformative impact of artificial intelligence in healthcare and underscores the importance of leveraging technology to address complex medical challenges.

Future Enhancements

Several avenues for future enhancements to the lung tumor detection project can further improve its functionality, accuracy, and usability:

Enhanced Model Architecture:

Explore more advanced CNN architectures or ensemble techniques to further improve the accuracy and robustness of tumor detection, considering factors such as feature extraction, model interpretability, and computational efficiency.

Incorporation of Clinical Data:

Integrate additional clinical data, such as patient demographics, medical history, and radiological findings, to enhance the predictive capabilities of the system and provide more personalized diagnostic recommendations.

Multi-Modality Imaging:

Extend the system to support multiple imaging modalities, including CT scans, MRI, and PET scans, to accommodate diverse clinical scenarios and provide comprehensive tumor detection across different imaging modalities.

Real-Time Image Analysis:

Implement real-time image analysis capabilities to enable instant feedback and decision support for healthcare professionals during medical imaging procedures, facilitating rapid diagnosis and treatment planning.

References:

[1] Ruchita Tekade,K. Rajeswari "Lung Cancer Detection and Classification Using Deep Learning" publisher-IEEE on 25 April 2019

[2] Bharathy S,Pavithra R,Akshaya. B "Lung Cancer Detection using Machine Learning" publisher-IEEE;Date Added to IEEE Xplore: 16 June 2022

[3] Susmita Das; Swanirbhar Majumder,"Lung Cancer Detection Using Deep Learning Network: A Comparative Analysis" publisher-IEEE;Date Added to IEEE Xplore: 21 December 2020

[4] Radhika P.R.; Rakhi A.S. Nair; Veena G."A Comparative Study of Lung Cancer Detection using Machine Learning Algorithms",publisher:IEEE;Date Added to IEEE Xplore: 17 October2019

[5] Varsha Prakash; P Smitha Vas,"Survey on Lung Cancer Detection Techniques",Publisher:IEEE,Date Added to IEEE Xplore: 18 September 2020

[6] Wasudeo Rahane; Himali Dalvi; Yamini Magar; Anjali Kalane; Satyajeet Jondhale;"Lung Cancer Detection Using Image Processing and Machine Learning HealthCare";Publisher:IEEE;Date Add to IEEEXplore: 29 November 2018

[7] Omar Khouadja; Mohamed Saber Naceur;"Lung Cancer Detection with Machine Learning and Deep Learning: A Narrative Review";Publisher:IEEE;Date Added to IEEE Xplore: 20 June 2023
[8] Pragya Chaturvedi, Anuj Jhamb, Meet Vanani and Varsha Nemade," Prediction and Classification of Lung Cancer Using Machine Learning Techniques" IOP Conf Ser., 2021 on SVM,KNN.
[9] Meraj Begum Shaikh Ismai "Lung Cancer Detection and Classification using Machine Learning Algorithm" Turkish Journal of Computer and Mathematics Education, 2021.
[10] R. L. Siegel, K. D. Miller, and A. Jemal, ``Cancer statistics, 2016", CA,Cancer J. Clin., vol. 66, no. 1, pp. 730,2016.
[11] Kanitkar SS, Thombare ND and Lokhande SS 2015 Detection of lung cancer using marker- controlled watershed transform Int. Conf. on Pervasive Computing (ICPC) pp. 1-6.
[12] Vas M and Dessai A 2017 Lung cancer detection system using lung CT image processing Int. Conf. on Computing, Communication, Control and Automation (ICCUBEA) pp. 1-5.
[13] Punithavathy K, Ramya MM and Poobal S 2015 Analysis of statistical texture features for automatic lung cancer detection in PET/CT images Int. Conf. on Robotics, Automation, Control and Embedded Systems (RACE) pp. 1-5.
[14] Sharma D and Jindal G 2011 Identifying lung cancer using image processing techniques Int. Conf. on Computational Techniques and Artificial Intelligence (ICCTAI) vol. 17 pp. 872-880.
[15] Asuntha A, and Srinivasan A 2020 Deep learning for lung Cancer detection and classification. Multimedia Tools and Applications pp 1-32.
[16] Teramoto A, Tsukamoto T, Kiriyama Y and Fujita H 2017 Automated classification of lung cancer types from cytological images using deep convolutional neural networks BioMed research International.