Lung Cancer Classification Using AI

M.Keerthana[1],,B.Surakshitha[2],B.Aravind[3],V.Rohith[4],Dr.M.Dhasaratham[5]
[5]*Professor, Department of Information Technology,TKR College of Engineering &Technology, Meerpet,Hyderabad*
[1,2,3,4] *UG Scholars,Department of Information Technology,TKR College of Engineering & Technology, Meerpet,Hyderabad*

***Abstract*—With the steady rise in lung cancer cases each year, early and precise diagnosis has become increasingly critical to ensure patients receive timely and effective treatment. To enhance diagnostic accuracy, low-dose computed tomography (CT) scans are widely used, offering detailed imaging that aids in detecting lung abnormalities. A key step in this diagnostic workflow is the identification of nodular formations within the lung tissue, as these nodules can be early indicators of malignancy. In this study, the LUNA16 dataset serves as the foundation for developing an automated detection system. This dataset includes 888 CT scans, each meticulously annotated with the exact coordinates of lung nodules. To prepare the data for analysis, three-dimensional volumesor cubes are extracted from each scan, centering the nodule within the volume. These volumetric samples provide rich spatial information essential for accurate detection. A 3D Convolutional Neural Network (3D CNN) is trained on these extracted cubes to learn the distinct features of lung nodules. This deep learning model effectively captures the three-dimensional structure of the nodules, allowing for robust recognition and localization across different scans.**

***Index Terms*—Lung Cancer Detection, Convolution Neural Network (CNN), VGG16, CT Scan Classification, Deep Learning,Medical Image Analysis.**

1. INTRODUCTION

Cancer remains one of the leading causes of death globally, with lung cancer standing out as the most fatal form. Each year, it is responsible for an estimated 2.1 million new cases and 1.8 million deaths worldwide. The key to improving patient survival lies in detecting lung cancer at its earliest stages, where treatment is more effective and the chances of recovery are significantly higher. Lung cancer develops when abnormal cells within the lung tissue multiply uncontrollably, eventually forming malignant tumors capable of invading nearby structures. The progression of the disease is classified into stages — with stages 1 and 2 confined to the lungs, while advanced stages indicate metastasis to surrounding organs and tissues. Today's diagnostic approaches rely on a combination of tissue biopsies and advanced imaging technologies such as computed tomography (CT) scans. Early identification through these methods is crucial, as it allows intervention before the cancer spreads beyond curable limits. Traditionally, computer-aided diagnosis has depended on image processing techniques to extract hand-crafted features from CT scans, enabling the differentiation between benign and malignant growths. However, crafting these features manually is not only labour-intensive but also requires deep domain expertise, making it a challenging task. To overcome these limitations, our project adopts deep learning techniques, specifically Convolutional Neural Networks (CNNs), which have the capability to automatically learn multi-level features directly from imaging data.

2. METHODOLOGY

i)Proposed Work

In this project, an advanced deep learning-based framework is proposed for the early detection and classification of lung cancer using CT scan images. The core objective is to develop an automated, accurate, and efficient system that assists medical professionals in diagnosing lung cancer at an early and treatable stage. The methodology employs the powerful VGG16 convolutional neural network (CNN) model, known for its superior performance in image classification tasks, and adapts it for medical imaging purposes. The approach combines image preprocessing, deep feature extraction, classification, and deployment into a complete diagnostic solution. The process begins with the collection of CT scan

images from publicly available and clinically validated datasets, including LUNA16 and the Data Science Bowl 2017 dataset. These datasets provide a rich source of annotated images where nodules are marked and classified. In preparation for model training, the CT scan images undergo preprocessing steps to ensure they are standardized and suitable for deep learning models. The images are resized to 224x224 pixels to match the input dimensions expected by the VGG16 model. Normalization is applied to scale pixel values, and data augmentation techniques such as rotation, flipping, and zooming are used to enhance the robustness of the model and reduce overfitting. Additionally, region of interest (ROI) masking is performed to isolate the lung areas from surrounding irrelevant tissues, improving detection accuracy.

Once the images are pre-processed, they are fed into the VGG16 model, which is employed as a feature extractor in this framework. VGG16, a 16-layer deep CNN architecture, uses multiple convolutional and pooling layers to automatically extract hierarchical features from the CT images. In this system, the pre-trained VGG16 model is fine-tuned with lung cancer data, allowing it to adapt its learned filters to the specific patterns found in cancerous and non-cancerous lung tissues. The deep features generated by the convolutional layers capture subtle structural differences that may not be visible to the naked eye, thereby enhancing diagnostic precision.

ii)System Architecture:

The system architecture is designed to efficiently handle the end-to-end process of lung cancer detection, from data ingestion to result visualization. It comprises several key modules: a data collection unit that gathers CT scan datasets; a preprocessing pipeline that standardizes images and isolates lung regions; a feature extraction module powered by the VGG16 model; a classification engine using customized dense layers; and a deployment interface built with Flask for real-time web access. The modular design ensures scalability, allowing for future enhancements such as the addition of more cancer types or real-time processing capabilities. Unified Modeling Language (UML) diagrams, including use-case, class, sequence, and activity diagrams, provide a clear blueprint of the system's structure and behaviour.

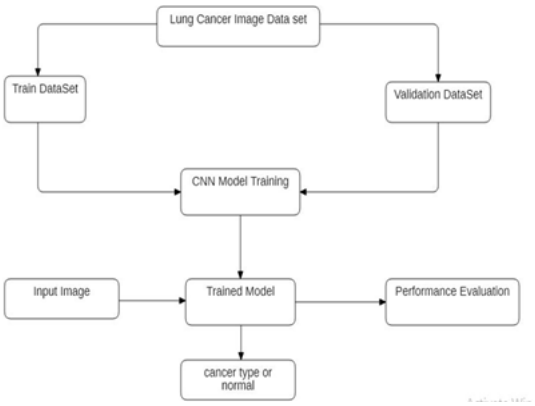

Fig 1 Proposed Architecture

iii)Data Collection

The success of any deep learning model largely depends on the quality and quantity of the data used for training. In this study, we utilize two benchmark datasets: the LUNA16 dataset, comprising 888 CT scans with annotated nodule locations, and the Data Science Bowl 2017 dataset, which includes 1,595 CT scans labeled for cancer classification. Data collection involves downloading these datasets, annotating the images using specialized labeling tools, and organizing them into training and testing sets. Each CT scan is carefully labeled to indicate the presence or absence of cancerous nodules, ensuring that the model learns from accurate and clinically relevant examples. An 80-20 train-test split is employed to evaluate the model's performance on unseen data, ensuring its generalization capabilities. By leveraging these diverse and well-annotated datasets, the system is trained to recognize a wide variety of lung cancer presentations.

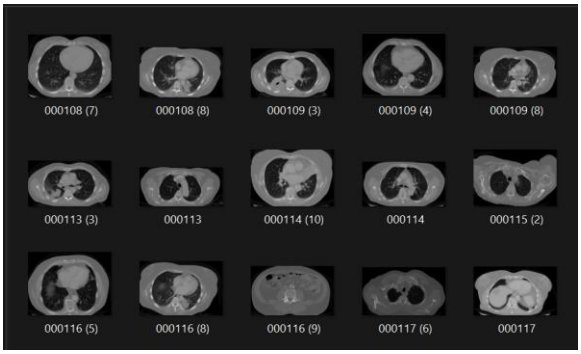

Fig 2 Classification Dataset

iv)Image Processing

Image pre processing is a critical step that significantly influences the performance of deep learning models.

In our system, CT scan images are first resized to a standardized resolution of 224x224 pixels, ensuring compatibility with the VGG16 input layer. Normalization techniques are applied to scale pixel intensities between 0 and 1, facilitating faster convergence during training. ROI masking is employed to extract lung regions from the scans, effectively reducing irrelevant background noise and focusing the model's attention on critical areas. Additional noise removal filters are applied to enhance the visibility of nodules, making it easier for the model to detect subtle patterns. These preprocessing steps not only improve the quality and consistency of the input data but also enhance the model's ability to learn and generalize from the training samples.

v)Data Augmentation

To further improve the robustness and generalization capability of the model, various data augmentation techniques are applied to the training dataset. Random rotations are introduced to make the model invariant to different orientations of CT scans. Flipping the images horizontally and vertically adds diversity to the dataset, while zooming and cropping simulate variations in image scale and focus. Brightness adjustments help the model handle images captured under varying conditions, ensuring it performs well across different scanners and settings. By artificially expanding the size of the training dataset through these augmentation methods, the model is exposed to a wider range of variations, reducing the risk of overfitting

vi) Algorithm

Convolutional Neural Network (CNN):

CNN is a neural network with multiple layers. It is a class of Deep Neural Network that is mostly used for the purpose of analysis of visual images. CNN contains one layer for input, called as input layer, next it has one or more middle layers which are called as convolutional layers, next the network consists of one or more fully connected layers and lastly the network is completed by adding an output layer in the respective order. We are using CNN as a building block for creating our models. It works as follows; a convoluted process is applied to the incoming data by the convolutional layers after which the outcome that is generated is passed to the next layer. This convolution layer has reaction similar to the human's neuron to vision process. Every neuron in the convolution uses information that it receives and processes the information that it is responsible for. Although there exist feed forward neural networks that can classify the data based on the features, it is not feasible to use such systems when it comes to processing pictures.

VGG16(Visual Geometry Group)

VGG16 is a deep convolutional neural network architecture developed by the Visual Geometry Group (VGG) at Oxford University. It gained popularity for its simplicity and high performance in image recognition tasks. The "16" in VGG16 refers to the 16 weight layers it contains including 13 convolutional layers and 3 fully connected layers. VGG16 uses very small convolutional filters (3x3) throughout the network, which allows it to capture fine details from input images while increasing depth and complexity.In this lung cancer detection project, VGG16 is used to analyze CT scan images of the lungs.

```
def train(epochs):
    print('Starting training..')
    for e in range(0, epochs):
        print('='*20)
        print(f'Starting epoch {e + 1}/{epochs}')
        print('='*20)

        train_loss = 0.
        val_loss = 0.

        vgg16.train() # set model to training phase

        for train_step, (images, labels) in enumerate(dl_train):
            optimizer.zero_grad()
            outputs = vgg16(images)
            loss = loss_fn(outputs, labels)
            loss.backward()
            optimizer.step()
            train_loss += loss.item()
            if train_step % 20 == 0:
                print('Evaluating at step', train_step)
```

Fig 3 VGG16

3. EXPERIMENTAL RESULTS

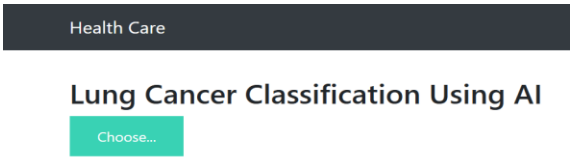

Fig 4 Upload input image

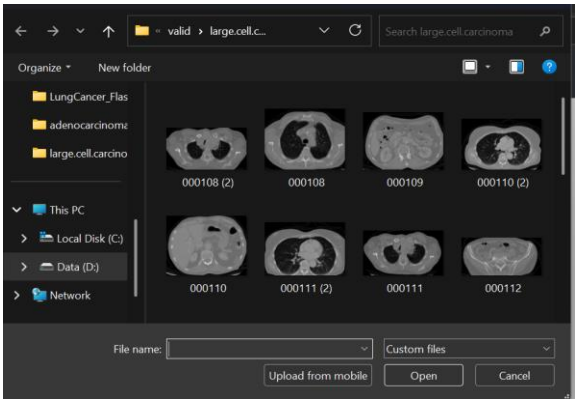

Fig 5 Input image folder

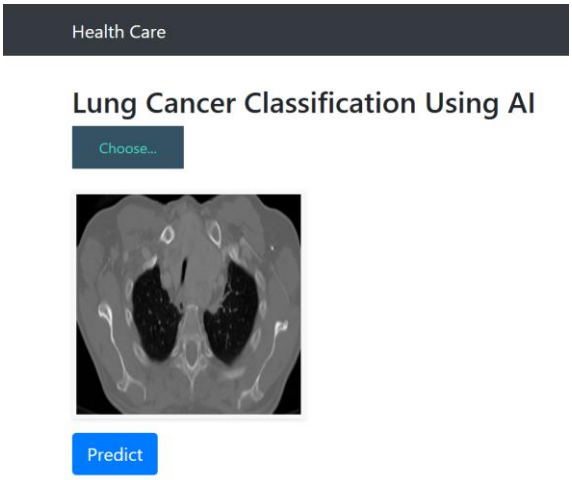

Fig 6 Uploaded Image

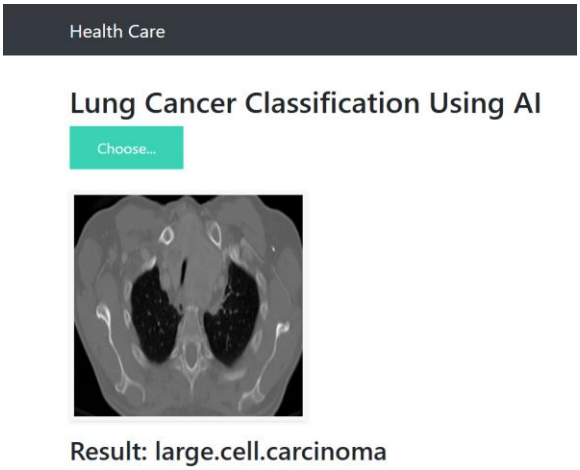

Fig 7 Predict result for given input

4. CONCLUSION

In conclusion, we developed an AI-driven system for the early detection and classification of lung cancer, leveraging deep learning techniques with the VGG16 architecture. By automating the analysis of CT scan images, the system aims to support clinicians in diagnosing lung cancer at an earlier stage, where treatment is more effective. Through advanced preprocessing, data augmentation, and transfer learning, the model achieved high precision and recall, demonstrating its reliability and practical applicability. The system's integration into a web-based platform further enhances its accessibility and ease of use in real-world clinical settings. Overall, this research highlights the transformative potential of AI in medical diagnostics, offering scalable and efficient solutions to improve patient outcomes.

5. FUTURE SCOPE

While the current system shows promising results, there are several avenues for future enhancement. Expanding the dataset with more diverse CT scans from different populations and equipment will improve the model's generalization and robustness. Implementing 3D CNN architectures can leverage the full volumetric data from CT scans, capturing richer spatial information and further improving detection accuracy. The system can also be extended to perform multi-class classification, distinguishing between various subtypes of lung cancer such as small cell and non-small cell carcinoma. Real-time deployment in radiology systems would enable instant diagnosis during patient scans, enhancing clinical workflows. Additionally, incorporating Explainable AI (XAI) techniques would allow the model to provide visual explanations for its predictions, increasing transparency and building trust among medical professionals.

REFERENCES

[1] G. Chartrand, P. M. Cheng, E. Vorontsov et al., "Deep learning: a primer for radiologists," RadioGraphics, vol. 37, no. 7, pp. 2113–2131, 2017.

[2] S. Wang, M. Zhou, Z. Liu et al., "Central focused convolutional neural networks: developing a data-driven model for lung nodule segmentation," Medical Image Analysis, vol. 40, pp. 172–183, 2017.

[3] A. A. A. Setio, F. Ciompi, G. Litjens et al., "Pulmonary nodule detection in CT images: false positive reduction using multiview convolutional networks," IEEE Transactions on Medical Imaging, vol. 35, no. 5, pp. 1160–1169, 2016.

[4] Q. Dou, H. Chen, L. Yu, J. Qin, and P. A. Heng, "Multilevel contextual 3-D CNNs for false positive reduction in pulmonary nodule detection," IEEE Transactions on Biomedical Engineering, vol. 64, pp. 1558–1567, 2017.

[5] F. Liao, M. Liang, Z. Li, X. Hu, and S. Song, "Evaluate the malignancy of pulmonary nodules using the 3D deep leaky noisy-or network," 2017, http://arxiv.org/abs/1711.08324.

[6] M. Kirienko, L. Cozzi, A. Rossi et al., "Ability of FDG PET and CT radiomics features to differentiate between primary and metastatic lung lesions," European Journal of Nuclear Medicine and Molecular Imaging, vol. 45, no. 10, pp. 1649–1660, 2018.

[7] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," in Proceedings of 3rd International Conference for Learning Representations, San Diego, CA, USA, December 2014.

[8] S. Edge, D. Byrd, C. Compton, A. Fritz, F. Greene, and A. Trotti, AJCC Cancer Staging Manual, Springer, Heidelberg, Germany, 7th edition, 2010.

[9] M. D. Hellmann, J. E. Chaft, W. N. William et al., "Pathological response after neoadjuvant chemotherapy in resectable non-small-cell lung cancers: proposal for the use of major pathological response as a surrogate endpoint," Lancet Oncology, vol. 15, no. 1, pp. e42–e50, 2014.

[10] F. O. Velez-Cubian, K. L. Rodriguez, M. R., au et al., "Efficacy of lymph node dissection during robotic-assisted lobectomy for non-small cell lung cancer: retrospective review of 159 consecutive cases," Journal of *oracic Disease, vol. 8, no. 9, pp. 2454–2463, 2016.

[11] N. C. Purandare and V. Rangarajan, "Imaging of lung cancer: implications on staging and management," Indian Journal of Radiology and Imaging, vol. 25, no. 2, pp. 109–120, 2015.

[12] Y. Wu, P. Li, H. Zhang et al., "Diagnostic value of fluorine 18 fluorodeoxyglucose positron emission tomography/computed tomography for the detection of metastases in non-small-cell lung cancer patients," International Journal of Cancer, vol. 132, no. 2, pp. E37–E47, 2013.

[13] H. Jin, Z. Li, R. Tong, and L. Lin, "A deep 3D residual CNN for false-positive reduction in pulmonary nodule detection," Medical Physics, vol. 45, no. 5, pp. 2097–2107, 2018.

[14] X. Zhao, L. Liu, S. Qi, Y. Teng, J. Li, and W. Qian, "Agile convolutional neural network for pulmonary nodule classification using CT images," International Journal of Computer Assisted Radiology and Surgery, vol. 13, no. 4, pp. 585–595, 2018.

[15] H.-H. Tseng, Y. Luo, S. Cui, J.-T. Chien, R. K. Ten Haken, and I. E. l. Naqa, "Deep reinforcement learning for automated radiation adaptation in lung cancer," Medical Physics, vol. 44, no.