Lung Cancer Detection Using Deep Learning

Aditi Singh[1], Juhi Singh[2], Raghuraj Singh[3], and Millan Saxena[4]
[1]*Corresponding author, Harcourt Butler Technical University, Kanpur-208002*
[2]*Contributing author, Harcourt Butler Technical University, Kanpur-208002*
[3]*Contributing author, Harcourt Butler Technical University, Kanpur-208002*
[4]*Contributing author, Founder, E-Cosys Consultancy Pvt. Ltd., New Delhi, 110077, Delhi*

Abstract: **Lung cancer is one of the major causes of death worldwide and its early and accurate diagnosis is crucial to improving survival rates. This paper focuses on the use of deep learning, such as Convolutional Neural Networks (CNN) to automatically classify lung cancer images as benign, malignant, or normal. We have trained a custom CNN model using the IQ-OTHNCCD dataset after applying pre-processing techniques to boost accuracy and reliability. The model performed exceptionally well, achieving 99.09% accuracy on the test set, with very low loss values. The results show that deep learning can effectively support the diagnostic process by making it faster and consistent. This AI-powered approach could help doctors catch lung cancer accurately at an earlier stage, potentially improving patient care and outcomes.**

Index Terms **AI in Healthcare, Benign vs. Malignant, Cancer Detection, CNN, Deep Learning.**

I. INTRODUCTION

In recent years, Convolutional Neural Networks (CNNs), has gained prominence as a powerful and promising approach [1] in medical imaging, offering automated solutions for disease detection and classification. A growing body of research has applied various CNN architectures, such as MobileNetV2, ResNet152V2, InceptionResNetV2, Xception, VGG-19, and InceptionV3 [3], to identify early stage lung cancer through the analysis of CT images. Among these, InceptionResNetV2 has achieved notably high detection accuracy (98.5%), while UNet has delivered strong segmentation performance (Jaccard index of 95.3%) [2]. Other studies have taken this further by developing advanced ensemble learning approaches, combining models such as ResNet152, DenseNet-169, and EfficientNet-B7, and introducing innovative weight assignment methods based on ROC-AUC and F1 scores. This has led to impressive results, with classification and sensitivity accuracies reaching 97.23% and 98.6% respectively [3].

Technology is rapidly changing the way we detect and diagnose lung cancer. Traditional imaging methods like chest X-rays, CT scans, and PET scans are essential tools for spotting problems in the lungs. But reading these images by hand can be time consuming and even the most skilled experts can sometimes miss early signs.

That's where image processing and deep learning are making a real difference. Image processing techniques help improve and refine medical images, making it easier to focus on areas that might look suspicious, such as lung nodules or tumors. At the same time, deep learning especially through models like convolutional neural networks (CNNs) can automatically scan and interpret these images, recognising patterns and helping distinguish between healthy and cancerous tissues with impressive accuracy.

By training on large numbers of medical images, these AI models are learning to assist radiologists in making quicker, more confident decisions. This could mean catching lung cancer earlier, starting treatment sooner, and ultimately improving outcomes for patients. As research moves forward, these smart technologies are becoming an increasingly valuable part of personalized lung cancer care.

These studies highlight the significant impact of deep learning in the early detection and classification of lung cancer, while also stressing the need for ongoing innovation and rigorous model validation to ensure effective real-world application.

II. LITERATURE REVIEW

Lung cancer is one of the most common and fatal diseases worldwide, contributing to the global number of cancer related deaths each year [2]. According to recent studies, over 200,000 new cases [6], [7] are diagnosed each year in the United States alone, making it a critical area of concern in medical research and healthcare [1]. Early detection plays very important role in improving the survival

rates[1], by as much as 50–75, as it allows for more effective treatment and a wider range of therapeutic options [3], [5]. However, early diagnosis continues to be a challenge due to the complex structure of lung nodules, overlapping tissue patterns, and the limitations of traditional diagnostic methods [4], [5]. Despite such encouraging progress, the implementation of deep learning systems in clinical practice still presents several hurdles. These include the need for model stability, accurate segmentation of nodules, reduction of false positives, and access to publicly available high-quality datasets [4]. Frameworks like DFCV have been proposed to focus on Data, Feature Selection, Classification Techniques, and View to guide the real-world implementation of DL-based systems [4]. The newer systems are now incorporating advanced 3D CT scan analysis, using networks such as UNETR for segmentation and self-supervised models for classification, reaching state of the art results, with a segmentation accuracy of 97.83% and a classification accuracy of 98.77% [6].

Several review articles have tried to map the progress of the field, but recent surveys offer a more comprehensive look, analyzing dozens of studies and summarizing the diverse models and outcomes achieved from 2016 to 2021 [1]. These reviews aim to help researchers to understand the strengths and limitations of various deep learning methods hence enhancing the decision making in lung cancer diagnosis. A detailed summary of the literature review giving dataset used, methodology applied, main findings, limitations and research gaps is shown in Table 1.

Table 1: Summary of the literature review

Research Paper Name	Dataset Used	Methodology	Main Findings	Limitation	Research Gaps
Deep learning applications for lung cancer diagnosis: A systematic review	LIDC-IDRI LUNA-16 NLST	The study reviewed 32 selected papers (2016–2021) on deep learning for lung cancer using a structured approach.	The paper reviews how deep learning, especially CNNs, helps in early lung cancer diagnosis with key models.	The review was limited by title-only, English searches, and deep learning models struggled with mixed, inconsistent data.	Not mentioned in paper
DFCV: a framework for evaluation deep learning in early detection and classification of lung cancer	LIDC-IDRI	The study used the DFCV framework to review 37 key deep learning papers on lung nodule classification.	The paper presents DFCV to help apply deep learning for lung nodule classification in clinics.	Unstable models, complex nodules, poor data, and lack of clinical focus.	Hard to apply in real clinics due to model and data challenges and it doesn't fully explain how well the model actually performs.
Lung cancer detection from thoracic CT scans using an ensemble of deep learning models	LIDC-IDRI	The study blends three models to classify lung nodules, weighting them based on performance.	The ensemble model achieved 97.23% accuracy and 98.6% sensitivity, reducing missed cases.	The model needs more testing, runs heavy, and lacks enough data—though synthetic images could help.	Spotting tiny lung nodules more accurately and combining models to make detection more reliable and accurate.
Comparative Analysis of Deep Learning Methods on CT Images for Lung Cancer Specification	Kaggle lung CT	The study used a Kaggle lung CT dataset, added augmentation, and trained with pre-trained deep learning models.	InceptionResNetV2 scored 98.5% in detection, InceptionUNet 95.3% in tumor mapping	The model needs better tumor typing, earlier detection, and combined features for stronger results.	Telling the difference between harmless and cancerous lung tumors
Medical Images Segmentation for Lung Cancer Diagnosis Based on Deep Learning Architectures	Decathlon	Used UNETR for lung scan segmentation and a self-supervised model for tumor classification.	The system hit 97.83% segmentation and 98.77% classification accuracy, outperforming 2D methods.	The model needs a strong GPU to work well, but using the cloud or a high-end computer can solve that.	Not mentioned
Intelligent deep learning algorithm for lung cancer detection and classification	LUNA-16	The study used CNNs on lung CTs, tested on LUNA16, and introduced the IDLA method.	IDLA uses CNNs to spot and classify lung cancer early, with high accuracy.	Smarter models can boost IDLA; early lung cancer is tricky, and AI can ease the load on doctors.	Not mentioned

III. METHODOLOGY

Complete methodology for the development of proposed model for effective and efficient early lung cancer detection is described in the following sub-sections.

A. Dataset and Class Labels

We have used the publicly available IQ-OTHNCCD lung cancer dataset [7], which consists of chest X-ray images labelled into three diagnostic categories shown in Fig. 1.

- Benign cases – non-cancerous lung conditions,
- Malignant cases –cancerous lung abnormalities,
- Normal cases – healthy individuals with no visible disease.

Each category was stored in a separate folder. This helped streamline data processing and label assignment during training.

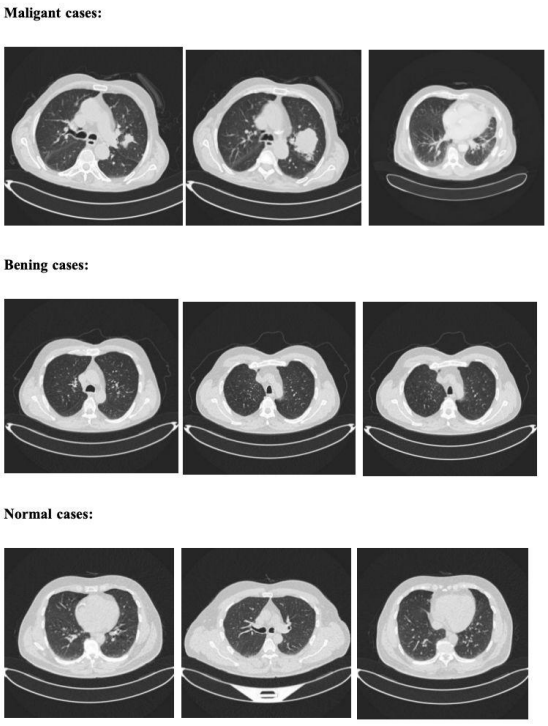

Fig. 1: Dataset from IQ-OTHNCCD lung cancer dataset [7]

B. Organizing the Dataset

To ensure balanced training and proper evaluation, the dataset is split into three parts:

- Training set (50%) – used in the model training.
- Validation set (25%) – was used to evaluate the model during training.
- Test set (25%) – kept completely unseen during training, used only for final evaluation.

Python's train test split function is used to randomly shuffle and divide the image paths while maintaining class balance. Checks to avoid empty folders or missing images are also added. Each subset (train, validation, test) is saved in its own directory structure organized by class. This made it easy to load images using TensorFlow's image loader functions.

C. Data Augmentation to Improve Learning

Since the dataset is relatively small, data augmentation is applied to generate more variety and help the model generalize better. Augmentation simulates real-world variations like different angles, lighting, and noise. The imgaug library is used to apply the following transformations:

- Rotation: Randomly rotated images by 90°, 180°, or 270°, mimicking different X-ray orientations.
- Flipping: Horizontally flipped 50% of the images to add symmetry.
- Translation and Scaling: Slightly shifted and zoomed images to simulate position or distance variations.
- Cropping and Padding: Added or removed pixels from the edges.
- Gaussian Noise: Added random noise to mimic imaging artifacts.
- Contrast Adjustment: Tweaked contrast to reflect machine-based differences in brightness.

Each original image was used to generate 4 new versions, increasing the effective dataset size significantly. We applied augmentation to both training and validation sets.

D. Pre-processing the images

Before providing images as input to the model, every image has been resized to 512 × 512 pixels for consistency. Images were then grouped into batches using TensorFlow's image dataset from directory() function.

To prepare the data for deep learning, values of the pixel have been normalized between -1 and 1. This scaling ensures smoother and faster learning, especially with ReLU activation and the Adam optimizer. To balance the performance and memory use, the batch size was set to 32.

E. Building the CNN Model

A customized Convolutional Neural Network (CNN) is developed using TensorFlow's Sequential API. The model is intentionally kept simple and efficient, yet deep enough to learn useful patterns from medical images. Breakdown of the architecture is given below.

- Conv2D layer (32 filters): Extracts basic spatial patterns like edges or corners.
- MaxPooling2D layer: Reduces image size and helps focus on the most relevant features.
- Second Conv2D layer (64 filters): Captures more complex patterns like textures or regions.
- Another MaxPooling2D: Further down samples the data.
- Flatten: Turns the 2D features into a 1D vector for dense layers.
- Dense layer (128 units): Learns high-level features for classification.
- Dense layer (64 units): Adds more depth to feature learning.
- Output layer (3 units, softmax): Predicts probabilities for each of the three classes.

ReLU activations have been used for hidden layers and softmax for the output, as it is ideal for multi-class classification.

F. Building the CNN Model

The model has been developed and trained using the following components.

- Optimizer: Adam, known for its adaptive learning and speed.
- Loss function: Sparse categorical cross entropy.
- Metric: Accuracy has been used for straightforward performance evaluation.
- Training: The model is trained over 10 epochs, with both training and validation accuracy and loss closely monitored throughout. Learning curves are plotted using Matplotlib to detect trends such as over-fitting or under-fitting.

IV. RESULTS AND DISCUSSION

The proposed CNN model performed remarkably and classified lung X-ray images as benign, malignant, or normal. As shown in Fig. 2 and Fig. 3 respectively, during training, it attains a remarkable accuracy of 99.43% with a very low loss of 0.0100, indicating that it learned the features effectively without simply memorizing the data. More encouraging is that even on the completely unseen test set, the model

maintains a high accuracy of 99.09%, with a test loss of just 0.0133, showing that it generalizes well beyond the training data.

A big part of these results come from the way data is prepared using smart pre-processing steps like normalizing pixel values and applying meaningful augmentations such as flipping, rotating, and adding noise to mimic real-world variations. Despite using a relatively simple, custom-built CNN, the model delivered strong results and could potentially be used in real clinical settings, especially in areas with limited access to expert radiologists. Moving forward, expanding the dataset and adding tools to explain the model's decisions (like visual heatmaps) would help make it even more reliable and trustworthy in real-world use.

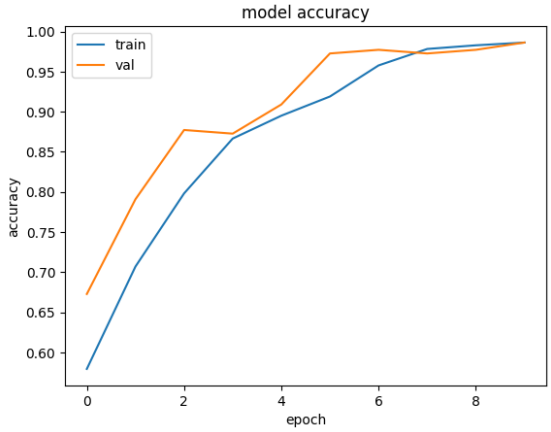

Fig. 2: The model reached near-perfect accuracy quickly

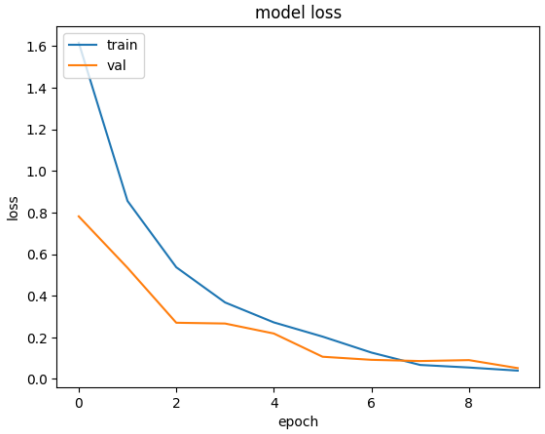

Fig.3: Rapid drop in training and validation loss, showing quick and effective learning

V. CONCLUSIONS AND FUTURE SCOPE

Deep learning shows powerful impact in enhancing the detection of lung cancer, especially through the use of CNN based models. Approaches using architectures like InceptionResNetV2, UNet, and EfficientNet-B7 have achieved impressive results, with detection and segmentation accuracies exceeding 97%. These models are proving to be powerful tools in analyzing CT scan images and supporting faster, more accurate diagnoses. At the same time, challenges such as high false positives, limited data availability, and the complexity of lung nodules remain ongoing concerns. To tackle these, researchers have introduced ensemble strategies and helpful frameworks like DFCV to guide model selection and improve real-world application.

Overall, while there's still work to be done before these systems are fully integrated into clinical practice, the progress so far is promising. With continued research and collaboration, deep learning can play a major role in making early lung cancer diagnosis, more accurate, and more life-saving. Building a custom deep CNN model specifically for lung cancer classification could lead to even better results than using general pre-trained models. Since lung CT scans have their own unique features, a model designed from the scratch can focus on what really matters in this context. By training the model from scratch and carefully adjusting things like the number of layers, filter sizes, and learning rate, important metrics like precision, recall, F1-score, and sensitivity can be improved. The challenges like noisy images and overlapping tissues can also be taken care of in future. With the help of data augmentation techniques, a custom model could become more reliable tool for lung cancer detection.

REFERENCES

[1] Seyed Hesamoddin Hosseini, Reza Monsefi, and Shabnam Shadroo, (2024), "Deep Learning Applications for Lung Cancer Diagnosis: A Systematic Review", Multimedia Tools and Applications, 83, 14305–14335, Springer, (69) https://doi.org/10.1007/s11042-023-16046-w.

[2] Muruvvet Kalkan, Mehmet S. Guzel, Fatih Ekinci, Ebru Akcapinar Sezer and Tunc Asuroglu, (2024), "Comparative analysis of deep learning methods on CT images for lung cancer specification", Cancers, 16, 3321, mdpi, (2), https://doi.org/10.3390/cancers16193321.

[3] Nandita Gautam, Abhishek Basu and Ram Sarkar, (2024), "Lung cancer detection from thoracic CT scans using an ensemble of deep learning models", Neural Computing and Applications, 36, 2459–2477, Springer, (12), https://doi.org/10.1007/s00521-023-09130-7.

[4] Abeer Alsadoon, Ghazi Al-Naymat, Ahmed Hamza Osman, Belal Alsinglawi, Majdi Maabreh and Md Rafiqul Islam, (2023), "DFCV: a framework for evaluation deep learning in early detection and classification of lung cancer" Multimedia Tools and Applications, 82, 44387–44430, Springer, (10), https://doi.org/10.1007/s11042-023-15238-8

[5] N. Sudhir Reddy, V. Khanaa, (2023), "Intelligent deep learning algorithm for lung cancer detection and classification" Bulletin of Electrical Engineering and Informatics, 12, 1747–1754, beei, (20), https://doi.org/10.11591/eei.v12i3.4579

[6] Yahia Said, Ahmed A. Alsheikhy, Tawfeeq Shawly and Husam Lahza, (2023), "Medical Images Segmentation for Lung Cancer Diagnosis Based on Deep Learning Architectures", Diagnostics, 13, 546, mdpi, (65), https://doi.org/10.3390/diagnostics13030546

[7] alyasriy, hamdalla; AL-Huseiny, Muayed (2023), "The IQ-OTH/NCCD lung cancer dataset", Mendeley Data, V4, doi: 10.17632/bhmdr45bh2.4