IJARSCT

International Journal of Advanced Research in Science, Communication and Technology

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

ISSN: 2581-9429

Volume 5, Issue 2, April 2025

Impact Factor: 7.67

Early Detection of Lung Cancer Using CNN: Enhancing Diagnostic Accuracy and Reducing False Positives

Fanta Jatta, Ebnesia Muchanga, Ei Su Po, Ravi Prakash Chaturvedi
Department of Computer Science and Application
Sharda University, Greater Noida, Uttar Pradesh, India
2022803173.fanta@ug.sharda.ac.in, 2022803102.ebnesiaines@ug.sharda.ac.in
2022811189.ei@ug.sharda.ac.in, rpchaturvedi51@gmail.com

Abstract: *This work aims to develop a deep learning application that can accurately perform early lung cancer diagnosis using Convolutional Neural Networks (CNN's). Using an image of the lungs that are scanned at high resolution helps improve accuracy of the diagnoses and aids in eliminating false behaviour that is common with traditional methods of diagnosis [2]. The main focus is to ensure that the necessary actions are taken as quickly as possible so as to increase the chances of survival. The proposed CNN-based system was fed with a set of 50 lung scans with high resolution and it achieved greater accuracy in identifying cancerous lesions than existing techniques with precise and recall metrics excelling [3]. In this way, this research seeks to assist radiologists by minimizing their diagnosis of patients while increasing the accuracy.*

Keywords: Convolutional Neural Networks

I. INTRODUCTION

1.1 Background

The late diagnosis of lung cancer has contributed to it being one of the top causes of cancer-related mortality. Biopsies and CT scans one of the many traditional diagnostic methods available, are often ignored due to their time-consuming nature and high false positive rates [7].

Problem Statement

Even though modern-day technology has been able to achieve several great milestones, the existence of false positives is still one of the major challenges in the early identification of lung cancer. The tackling of reducing false positives and maintaining sensitivity has been stated as key in timely treatment of the disease [4]. This discovery has resulted in the nurturing of research in tools which are more accurate and accessible through the assistance of AI.

1.2 Objectives

With the fueling advancements in technology, the following objectives have been set with the help of a CNN-based approach for:

- The development of deep learning algorithms which improve diagnostic accuracy.
- The implementation of high resolution image processing to minimize false positive rates.
- The result being the enhanced and quicker identification of cancer.

II. RELATED WORK

Many works have investigated deep learning models for detection of cancer. Lakide & Ganesan [8] proved that CNN could usurp performing traditional methods in the blame game of identifying lung cancer even during its early stages. Likewise, Kanagalakshmi & Nisha

DOI: 10.48175/IJARSCT-24944

[5] used graph-based CNN approach and were able to improve the accuracy of detection. Works including Heyat et al. [3], and Aelgani et al. [4] showed the need for reducing the false positive rate by redesigning the structures of models and using other more developed feature extraction methods.

III. VISUALIZATION OF DATASET

Visualization of the dataset is a key component that assists in understanding the data facilitating the model to perform better. However, CT scan images vex to be visualized on normal desktops or browsers because of the specialized format to which they are in. As a solution towards this problem, there is the use of Pydicom library to read and visualize the DICOM format Pydicom Library - DICOM files into image arrays to better visualize the lung CT scan images of interest, and also aids them in extracting metadata embedded in CT images including:

Pydicom Library: This library enables the conversion of DICOM H6 files into arrays so as to better visualize the CT scan images of the lungs in question. It also aids in extracting embedded metadata in the CT images who includes:

- Patient's Name
- Patient's ID
- Date of Birth
- Position of the Image
- Number of the Image
- Doctor's Name and Date of Birth

LUNA16 Dataset Structure

- Various patient IDs have been used as names for the subdirectories in which the LUNA16 dataset is kept.
- The LUNA16 dataset has a patient's subfolder which has close to 180 2D images, when increased, gives a 3D lung image. [6]

The amalgamation of such slices can be bundled to generate a detailed image of what lungs appear like enabling the viewers to identify potential cancer cells and other distinguishing features.

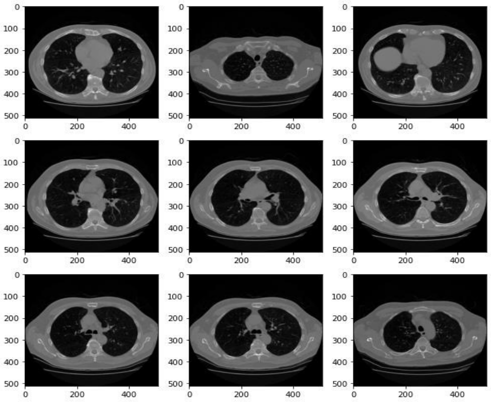


Fig. 1: Original DICOM slices from LUNA16 dataset

3D Image Creation

By compositing all the 2D layers from one of the subfolders, a 3D representation of the lung is generated that is quite useful in determining the volumes of cancer lesions for lung disorders.

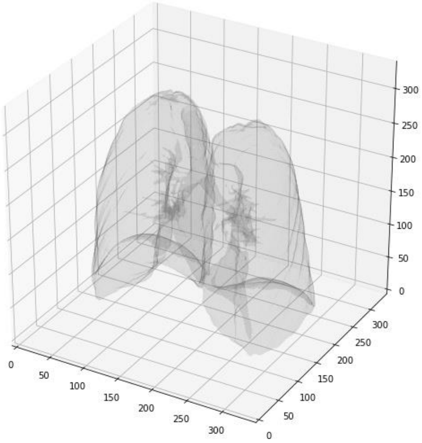


Fig. 2: 3D image of lungs of a single patient

IV. PROPOSED MODELS

The proposed model has utilized CT scan images in the locating of lung cancerous tissues through a convolutional neural network technique. The tasks include:

LUNA16 Data Set Preprocessing

This step involves Resizing, Normalizing the data to achieve same image dimension as well as augmentation.

Lung Segmentation Implemented using Watershed Algorithm

The watershed algorithm detects the lung area, and segments it into binary masks using the semantic segmentation method to expose the lung area to further analysis.

CNN Models Used

For several classification tasks over various datasets, three convolutional neural networks were used to evaluate the performance of several convolutional neural networks over CT datasets:

The overview of models constructed is presented in the following table:

- Model 1: Sequential_1 : It is a basic model based on a CNN that contains dense and max pooling layers along with conv layers and dropout layers. This model performed satisfactorily for simple classification tasks [10].
- Model 2 : Sequential_2 is also a deeper model which has performed well on multiple datasets and has multiple convolutional layers, fully connected layers and max pooling layers due to its depth [12].
- Model 3: Transfer Learning with VGG-16 : This model adjusted the last three fully connected layers of the VGG-16 architecture as they were concerned with binary classification (cancerous vs non-cancerous types). It obtained reasonably high accuracy thanks to the pre-trained weights on the ImageNet dataset and the fine-tuning with the specific dataset used our model [14].

Training Information

1. Data Preparation

The watershed algorithm was employed to create binary lung segmented masks.

Shapes of images:

(512, 512, 1) for Sequential_1 and Sequential_2.

(512, 512, 3) for VGG-16

2. Training Configuration

For all models, a batch size of 32 was used.

Epochs:

30 epochs with 100 images each for Sequential_1 and Sequential_2

30 epochs with 500 images each for VGG-16.

These data augmentation techniques were implemented

Shear range, Zoom range, Horizontal flip, Rotation range, center shift.

3. Binary Classification Setup

The last layer binary classification has one node and this one has a sigmoid activation function.

4. Model Checkpointing

The best model as assessed in terms of accuracy on the validation set was saved by utilizing the TensorFlow Keras callbacks.

An entire training session of 50 epochs was performed to formulate comparison graphs.

For all models, a batch size of 32 was used.

Epochs:

30 epochs with 100 images each for Sequential_1 and Sequential_2.

30 epochs with 500 images each for VGG-16.

These data augmentation techniques were implemented

Shear range, Zoom range, Horizontal flip, Rotation range, center shift.

5. Binary Classification Setup

The last layer binary classification has one node and this one has a sigmoid activation function.

6. Model Checkpointing

The best model as assessed in terms of accuracy on the validation set was saved by utilizing the TensorFlow Keras callbacks.

An entire training session of 50 epochs was performed to formulate comparison graphs.

V. METHODOLOGY

5.1 Dataset Collection

- A Certain aggregated dataset of lung scans conducted on 50 patients was utilized for training as well as validation.
- According to radiologists, every scan was annotated as either containing cancer or not hence providing quality annotations (Jariya & Jain, 2024).
- There was uniformity with the size of images and normalization of pixel values in pre-processing as this allowed the convergence training to be faster. Moreover, data augmentation techniques, such as flipping, rotation, and scaling were leveraged to improve the generality of the model.

5.2 Model Architecture

The layers of the CNN model comprised:

- Convolutional Layers: feature extraction from the images surface textures.
- Pooling Layers: Minimized model complexity while conserving necessary information.
- Fully Connected Layers: Were purposed for translating the synthesized high-level features into the cancerous or non-cancerous classes.

DOI: 10.48175/IJARSCT-24944

- Dropout Layers: Randomly turned off some neurons during the training in order to combat overtraining (Lakide & Ganesan, 2024).

5.3 Training Process

- Due to lack of GPU compatibility during this research, the model was trained on a CPU environment using Python 3.9 and TensorFlow.
- The objective function utilized here was cross-entropy loss and the Adam optimizer was chosen with a learning rate of 0.001.
- In the current study, training was done in 50 epochs with a batch size of 32. An approach of early stopping was applied to end the training when no more improvement was seen in the validation loss.

5.4 Validation and Testing

- 20 percent of the entire dataset was set aside to carry out validation.
- Various performance metrics such as accuracy, precision, recall, F1- score and ROC-AUC were used to gauge the effective of the model.

VI. TRANSFER LEARNING: VGG16-NET

VGG Net is one of the most famous Convolutional neural networks which is developed by Simonyan and Zisserman from visual geometry Group VGG University of Oxford in 2014 and came second in the ILSVRC (ImageNet large scale visual recognition competition) 2014. It was trained on the ImageNet ILSVRC dataset containing

1.3 million training images, 100,000 test images and 50,000 validation images, achieving a validation accuracy of 92.7% on test set. Furthermore, the networks' ability to work in quantitative applications such as estimating heart rate from body motion and detecting pavement distress is a testimony of its capabilities in object identification and feature recognition.

VGG16-Net Architecture

The VGG Net is designed to improve classification performance by deepening the architecture of CNNS. The two major popular types are the VGG16 and VGG19, which have 16 and 19 weight layers respectively. This VGG Net takes as input RGB images of size 224×224, and it does so by creating multiple layers of 3×3 convolutions with a stride of 1 followed by five layers of max-pooling to help reduce the dimension. Following the convolution stack, there are three fully connected layers of 4096, 4096 and 1000 channels respectively and these are followed by a SoftMax layer which is responsible for classification.

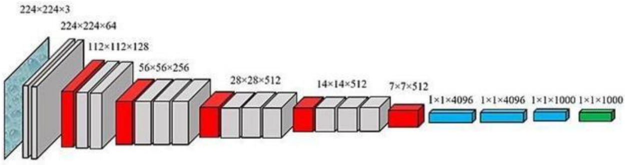


Fig. 6: VGG16-Net Working Architecture

Configuring VGG16 to Fit Our Dataset Because our dataset consists of lung images of 512 × 512 high resolution, we had to modify VGG16 so that the input layer can accept images this size, hence the change in the model. The modified model was compiled with the following properties:

- Optimizer: Adam
- Learning Rate: 0.0001
- Loss Function: Binary Cross entropy
- Metrics: Accuracy

This model utilizes VGG16's learned weights while retraining this network model to achieve higher accuracy for our classification task (cancerous vs. non-cancerous lung scans).

VII. CONCLUSIONS AND RESULTS

The following models are shows accuracy and loss with respect to graph in our project.

Model: Sequential_1

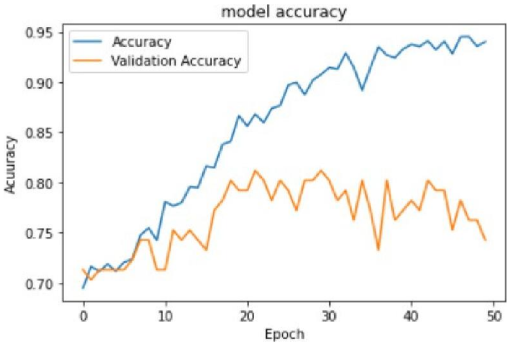
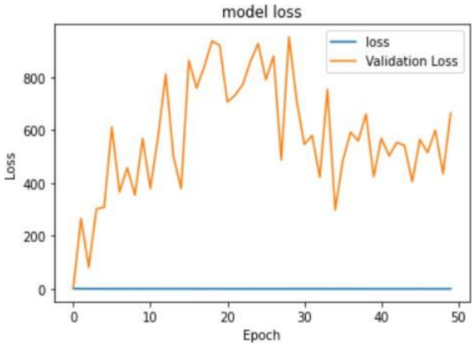


Fig 7 . Training and Validation Accuracy of model Sequential_1



Fg 8. Training and Validation loss of model Sequential_1

Model: Sequential_2

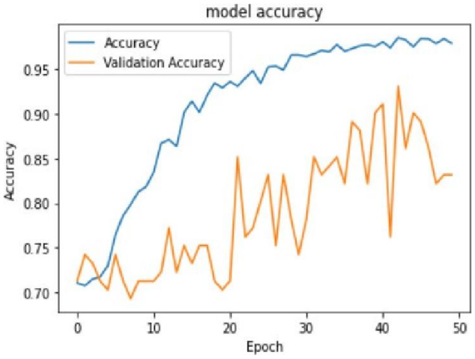


Fig 9 . Training and Validation Accuracy of model Sequential_2

DOI: 10.48175/IJARSCT-24944

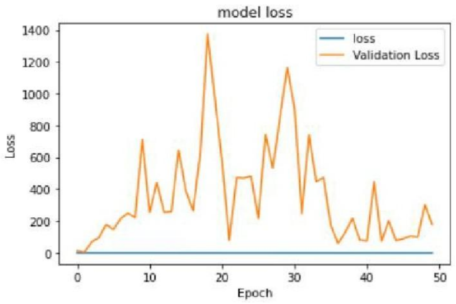


Fig 10. Training and Validation loss of model Sequential_2

Model: VGG_16

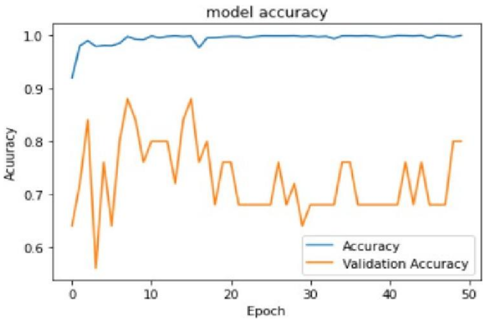


Fig 11. Training and Validation Accuracy of model VGG_16

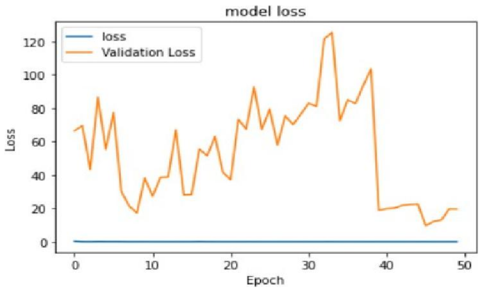


Fig 12. Training and Validation loss of model VGG_16

Tabular Comparison of model training accuracy, model loss, tested model accuracy and model loss values:

Index	Model	Train Acc.	Train loss	Val. Acc.	Val. loss
1.	Sequential_ 1	90.77%	0.2242	81.19 %	712.0875
2.	Sequential_ 2	98.53 %	0.0442	93.07%	74.5244
3.	VGG_16	99.84 %	0.0046	88.00 %	28.2614

VIII. FUTURE WORK

- Datasets Expansion: Implement larger and more complex datasets to enhance generalization of the model [6].
- Transfer Learning: Apply higher efficient models that are previously trained, for example ResNet, EfficientNet etc. [15].
- Real Time Applications: Create a web or mobile application that would allow users to receive real- time lung cancer diagnosis.

- Clinical Trials: Partner with medical institutions to perform real checks [1].

REFERENCES

[1]. Saxena, S., Prasad, S. N., & Polnaya, A. M. (2025). *Hybrid deep convolution model for lung cancer detection with transfer learning*.

[2]. Molaeian, H., Karamjani, K., & Teimouri, S. (2024). *The potential of convolutional neural networks for cancer detection*.

[3]. Heyat, M. B. B., Ansari, M. M., & Ullah, H. (2025). *SVMVGGNet-16: A novel machine and deep learning-based approach for lung cancer detection using combined SVM and VGGNet-16*.

[4]. Kanagalakshmi, K., & Nisha, H. B. (2024). *Lung cancer prediction with improved graph convolutional neural networks*. SSRN Electronic Journal.

[5]. Jariya, N., & Jain, M. (2024). *Lung sound-based disease detection with deep learning*. International Journal of Trend Research in Machine Learning, 11(11), 45-58.

[6]. Okae, P., Addo, T., Owusu-Afari, J. B., & Bondzie, G. (2025). *Cancer detection and classification using CNN model*. Research

DOI: 10.48175/IJARSCT-24944