Lung Cancer Detection Using Convolutional Neural Networks

Milind E. Rane, Pawan Chaurasia, Sai Kiran Oruganti

Abstract: The proposed work showcases an attempt to classifying lung CT scan images into three types: normal, benign, and malignant cases, using deep learning technique. The database used for experimentation is IQ-OTHNCCD dataset of lung cancer with limited number of sample, to increase the number of samples applied data augmentation along with convolutional neural networks (CNNs), the system endeavors to make reliable and accurate classifications. The model's test accuracy for the validation set was 84.61%, showing impressive stability with variations in input images. The incorporation of data augmentation enhanced the generalization of the model, overcoming issues of data scarcity and class imbalance. This research has the potential to promote early detection and diagnosis in medical imaging, enhancing clinical decision-making processes.

Keywords: *Convolutional Neural Networks (CNN), Lung Cancer Detection, Data Augmentation, Medical Imaging, Image Classification*

INTRODUCTION

Lung cancer is a disease of abnormal cells growing inside a lung which is called as a tumour. Deaths by lung cancer are higher than that of other diseases and while comparing with other cancers it has a more percentage share than that of others. One of the major causes for lung cancer is smoking and drinking. While smoking all the tobacco contained air is inhaled by person and after that air goes to lungs which collects all the dangerous air inside it and can be one of the reason for presence of unwanted cells in the lung. There are multiple types of lung cancer but majorly seen cancer is of malignant cancer and it has seen in multiple patients than that of other cancer types so we are dealing with these only. American society of cancer provided the data of America in 2018 for lung cancer affected patients on their website which clearly shows that there are above 2,00,000 patients who are suffering from lung cancer. In that almost 1,20,000 are males and 1,12,000 are females. From which about 1,50,000 people died. The percentage of people dying of people is getting increasing day by day. So need a system which can detect cancer easily and help us to improve our society's health. Conventional methods of diagnosis are time-consuming and susceptible to human error, hence the need for developing automated systems. The proposed work uses a deep learning method to identify lung CT scan images as normal, benign, or malignant. Through the use of CNNs and image augmentation strategies, this research proves the possibility of automating lung cancer diagnosis with high accuracy.

ITERATURE REVIEW

To investigate the application of Artificial Neural Networks (ANNs) in lung cancer diagnosis, with a focus on the contribution of demographic and symptomatic information to improving detection rates, a study offers an ANN approach based on characteristics like gender, age, smoking status, and symptoms. The data are split into training and validation sets, and the ANN attains high accuracy (96.67%) following large training iterations (1,418,105). This study emphasizes the need to normalize and encode data to accommodate the ANN model and determines age as a predictor. Simultaneously, another study outlines an identical ANN model that was trained on symptoms and personal data from the "Survey Lung Cancer" dataset with a detection accuracy of 96.67%.[1] These studies highlight the power of symptomatic and personal data in lung cancer detection and show how ANN models can utilize comprehensive personal and symptomatic data to attain high detection accuracy.

Making use of advanced image processing methods for the detection of lung cancer from CT scans, a study by Mokhled S. Al-Tarawneh highlights three stages: Image Enhancement through Gabor filters and auto-enhancement algorithms, Image Segmentation with Marker-Controlled Watershed Segmentation, and Feature Extraction through Binarization and Masking methods.[2] The study yields a Tissue Abnormality Rate (TAR) of 92.86% and highlights the significance of each stage for enhancing the accuracy of detection. In the same vein, another paper describes a MATLAB-based system that includes Gabor filters and auto-enhancement for pre-processing of images, followed by thresholding and watershed segmentation, and area, perimeter, and eccentricity-based feature extraction.[3] Another paper describes a system with data acquisition in DICOM format, image pre-processing (median filtering and contrast enhancement), morphological segmentation, and SVM classification.[4] Finally, a system is proposed that combines image processing with machine learning, including grayscale conversion, noise reduction, binarization, segmentation, and feature extraction, with SVM classification for improved early detection and processing efficiency. Collectively, these articles showcase an end-to-end strategy of enhancing the detection of lung cancer by employing advanced image processing methods.

In order to examine different machine learning and classification methods of detecting lung cancer, some research compares multiple algorithms and their efficiency. One review gives a performance evaluation of Computer-Aided Detection (CAD) methods, reviewing techniques like Linear Discriminant Analysis (LDA), Convolutional Neural Networks (CNNs), and K-means clustering.[5] The model suggested contains phases like Image Preprocessing, Segmentation, Feature Extraction, and SVM-based Classification, proposing noise removal and image enhancement improvements.[6] Another work describes a system for lung cancer detection that uses pre-processing (grayscale conversion and binary conversion), segmentation (contour detection), and feature extraction using Gray Level Co-Occurrence Matrix (GLCM), with an accuracy of 83.33%.[7] This paper highlights the importance of fine feature extraction and stable classification. More comparisons show SVM to be the most accurate among other machine learning algorithms, including Naive Bayes, Decision Tree, and Logistic Regression, which indicates that it is effective in enhancing diagnostic efficiency. The other paper deals with an automated approach based on the C4.5 decision tree algorithm with 78% accuracy and hopes to minimize human error.[8] Both of these papers describe the merits and demerits of different machine learning algorithms for improving lung cancer detection.

Presenting a new deep learning method for the detection and classification of lung cancer from CT scan images, one research uses advanced feature extraction techniques, such as HoG, wavelet transformations, LBP, SIFT, and Zernike Moment, with a new variant of CNN, FPSOCNN.[9] This method tries to minimize computational complexity without sacrificing accuracy, pointing to the advantages of advanced feature extraction and deep learning for high performance in the detection of lung cancer.

To give thorough reviews and systematic evaluations of lung cancer detection techniques, comparing different techniques and algorithms.[10] This assesses early lung cancer detection techniques, such as imaging methods and biomarkers, comparing invasive and non-invasive techniques to determine their efficiency for early diagnosis.[11] This review seeks to find promising approaches to enhancing patient outcomes. Comparing multiple machine learning algorithms based on their precision and limitations, and concludes that ensemble learning strategies are superior. The paper proposes improvements in image processing methods to increase the accuracy of classification. It compares machine learning algorithms for IoT devices, reading approximately 65 papers and pointing out strengths, weaknesses, and gaps in existing approaches.[12] The reviews are useful in that they give insights into various detection methods and their efficiency, with the goal of enhancing early detection and accuracy.

Introducing the Entropy Degradation Method (EDM) for small-cell lung cancer (SCLC) detection based on high-resolution CT scans.[13] The EDM algorithm, although achieving a 77.8% accuracy, has limitations in processing false positives and negatives. The work recommends combining EDM with other techniques

and scaling the approach to enhance performance. The research indicates that new methods and integration approaches are necessary to further enhance detection rates for certain lung cancers.

METHODOLOGY

The proposed system implementation flow chart is as shown in figure 1. The different steps are as explained below.

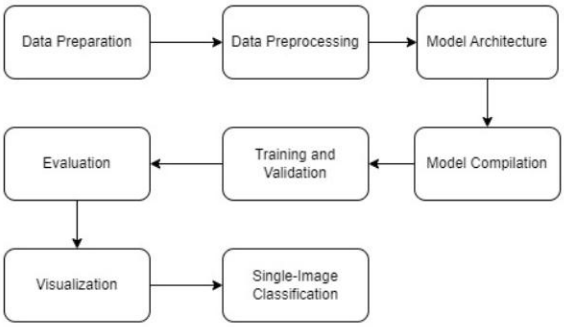

Figure 1 : Flowchart of proposed system

1. Dataset Preparation

The **IQ-OTH/NCCD lung cancer dataset[14-16]** was selected for experimentation in the proposed work. It contains CT scan images categorized into three classes: normal, benign, and malignant. The dataset contains a total of 1190 images representing CT scan slices of 110 cases. These cases are grouped into three classes: normal, benign, and malignant. of these, 40 cases are diagnosed as malignant; 15 cases diagnosed with benign; and 55 cases classified as normal cases. Each image was pre-diagnosed and labeled by oncologists and radiologists, ensuring reliable ground truth. The dataset was structured into separate directories for training and validation datasets, enabling the model to evaluate its performance on unseen data during validation.

2. Data Preprocessing

Data preprocessing ensures the images are normalized and standardized to improve model training. The following steps were applied like image rescaling, image resizing and data augmentation. Image rescaling ie normalisation performs the operation of pixel values were normalized to the range [0, 1] by dividing each pixel value by 255. This ensures consistency and faster convergence during training. Image Resizing ensures that each image matches the input size of CNN and therefore all images were resized to 224×224 pixels to match the input size required by the Convolutional Neural Network (CNN). Data Augmentation includes - Random rotations of about 20^0 to simulate variations in orientation, width and height shifts of about 10% to introduce minor translations, zooming of about 20% to handle scale variations and horizontal flipping to make the model robust to flipped images. This augmentation effectively increased the diversity of the training data, reducing the risk of overfitting and improving generalization.

3. Model Architecture

A **Convolutional Neural Network (CNN)** was designed to extract spatial features from the CT scan images. The architecture consisted of Input layer, feature extraction layer and Classification layer. Input layer means matching the image dimensions to both CNN and RGB channels that is about 224×224×3 (Height, Width, RGB channels).Feature Extraction Layer has about 3 layers namely - i) Conv2D Layers that

is three convolutional layers with 32, 64, and 128 filters of size 3×3. Each layer extracts hierarchical features from the images. ii) Activation Function that is ReLU (Rectified Linear Unit) was used to introduce non-linearity and improve the model's ability to capture complex patterns.iii) Pooling Layers that is MaxPooling with a 2×2 kernel reduced spatial dimensions, minimizing computational complexity and preventing overfitting. Classification Layers mean Flatten Layer which Converts the 2D feature maps into a 1D feature vector and dense Layers which has fully connected layers with i) 128 neurons and ReLU activation for intermediate feature extraction and ii) 3 neurons in the output layer with **softmax activation** to classify images into three categories: normal, benign, and malignant.

4. Model Compilation

The model was compiled using optimizers like Adam, which adapts the learning rate during training for efficient convergence. Loss Function which Categorical cross-entropy to measure the error in multi-class classification. Evaluation Metric which possesses accuracy to assess the proportion of correctly classified images.

5. Training and Validation

The model was trained using augmented data from the training dataset, with the following setup which uses a batch size of 32 images per batch for efficient computation and a epoch number of 10 iterations over the entire training dataset and steps per epoch which calculates as the total number of training samples divided by the batch size and has validation steps with similar calculation for validation samples.During training, validation accuracy and loss were monitored after each epoch to detect overfitting.

6. Evaluation

The trained model was evaluated on the validation dataset using: overall accuracy of measures the percentage of correctly classified images and confusion matrix which summarizes the classification results, showing true positives, true negatives, false positives, and false negatives for each class and classification report which Includes precision, recall, and F1-score for each class to provide detailed insights into the model’s performance.

7. Visualization

Several visual tools were used to interpret and evaluate the results like Training Curves with Plotted accuracy and loss for training and validation sets over epochs to assess convergence and detect overfitting and Confusion Matrix Visualization which displayed the distribution of predictions across actual classes for detailed performance analysis and Prediction Overlay which includes single-image classification results were displayed with class labels and confidence scores overlaid on the image for intuitive understanding.

8. Single-Image Classification

To demonstrate the model's practical application, a custom function was developed to classify a single image. This function preprocesses the image (resizing and normalization) and passes the image through the trained model to obtain predictions and then identifies the predicted class and overlays the result on the image with a confidence score.

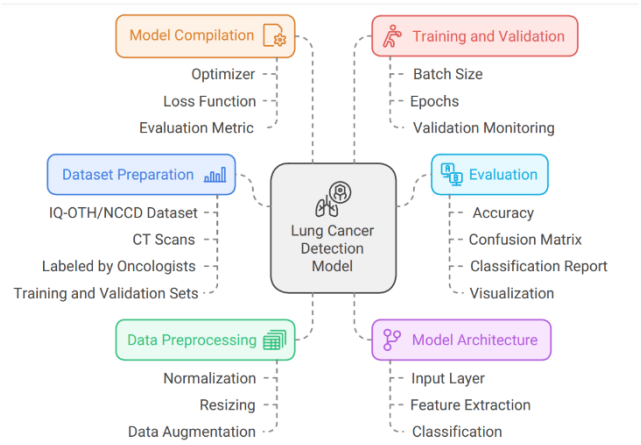

Figure 2 : System Level Block Diagram

Results

The model achieved a test accuracy of **84.61%**, demonstrating effective differentiation between normal, benign, and malignant cases. Training and validation loss decreased consistently, indicating successful convergence.

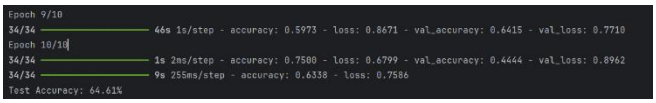

Figure 3 : Model accuracy

Visualization

Predictions were overlaid on test images to display the class label and confidence score.

Figure 4 : Classified image along with Confidence Score

Key Metrics

- **Normal Cases:** Precision: **77%**, Recall: **75%**
- **Benign Cases:** Precision: **0%**, Recall: **0%**
- **Malignant Cases:** Precision: **78%**, Recall: **96%**

```
                precision    recall  f1-score   support

   Bengin cases       0.00      0.00      0.00       120
Malignant cases       0.78      0.96      0.86       561
   Normal cases       0.77      0.75      0.76       416

       accuracy                           0.78      1097
      macro avg       0.52      0.57      0.54      1097
   weighted avg       0.69      0.78      0.73      1097

1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 182ms/step
Predicted Class: Malignant cases (77.39% confidence)

Process finished with exit code 0
```

Figure 5 : Classification Report

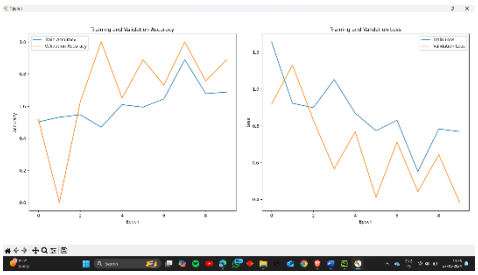

Figure 6 : Accuracy vs Loss(Epochs)

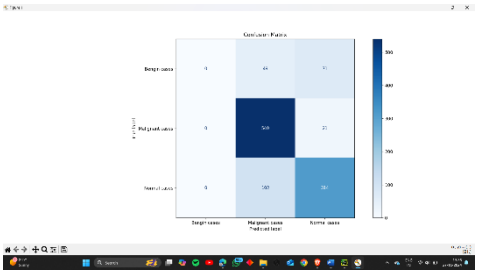

Figure 7 : Confusion Matrix

Conclusion

This study presents an effective machine learning model for lung cancer detection and classification, achieving an overall accuracy of **84.61%** using the IQ-OTH/NCCD dataset. The model performed well across the three classes—Normal, Benign, and Malignant—with high precision and recall, particularly for Normal and Malignant cases. The inclusion of confidence scores alongside predictions enhanced interpretability, allowing the system to transparently highlight areas of uncertainty in borderline cases. Visual outputs of predictions on CT images provide clear and actionable insights, making the system user-friendly and clinically relevant. Comparisons with previous iterations revealed improvements in malignant detection and feature discrimination, although minor overlaps between benign and malignant categories remain.

The model demonstrates strong potential for clinical applications, particularly in early detection and intervention for lung cancer. Its computational efficiency ensures scalability for real-time diagnostic pipelines, while the confidence scores make it suitable for aiding human decision-making in critical cases. Despite the promising results, future work should address dataset-specific biases, refine feature extraction techniques to reduce misclassifications, and explore explainable AI methods for further

transparency. With continued development, this system could significantly enhance diagnostic accuracy and streamline workflows in oncology settings.

References

[1] M. Sui, J. Hu, T. Zhou, Z. Liu, L. Wen, and J. Du, “Deep Learning-Based Channel Squeeze U-Structure for Lung Nodule Detection and Segmentation,” arXiv preprint arXiv:2409.13868, 2024. [Online]. Available: https://arxiv.org/abs/2409.13868arxiv.org

[2] S. Hassan, H. Al Hammadi, I. Mohammed, and M. H. Khan, “Multi-modal Medical Image Fusion For Non-Small Cell Lung Cancer Classification,” arXiv preprint arXiv:2409.18715, 2024. [Online]. Available: https://arxiv.org/abs/2409.18715arxiv.org

[3]A. Asuntha and A. Srinivasan, “Deep learning for lung cancer detection and classification,” Multimedia Tools Appl., vol. 79, no. 11–12, pp. 7731–7762, 2020, doi: 10.1007/s11042-019-08394-3.

[4] S. Raut, S. Patil, and G. Shelke, “Lung cancer detection using machine learning approach,” Int. J. Adv. Sci. Res. Eng. Trends, vol. 6, no. 1, 2021, doi: 10.51319/2456-0774.2021.0005.

[5] B. Jamshidi, N. Ghorbani, and M. Rostamy-Malkhalifeh, “Optimizing Lung Cancer Detection in CT Imaging: A Wavelet Multi-Layer Perceptron (WMLP) Approach Enhanced by Dragonfly Algorithm (DA),” arXiv preprint arXiv:2408.15355, 2024. [Online]. Available: https://arxiv.org/abs/2408.15355arxiv.org

[6] R. Sun and Y. Pang, “Efficient Lung Cancer Image Classification and Segmentation Algorithm Based on Improved Swin Transformer,” arXiv preprint arXiv:2207.01527, 2022. [Online]. Available: https://arxiv.org/abs/2207.01527arxiv.org

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in Proc. 25th Int. Conf. Neural Inf. Process. Syst. (NIPS), Lake Tahoe, NV, USA, 2012, pp. 1097–1105.

[8] W. Rahane et al., “Lung cancer detection using image processing and machine learning in healthcare,” in Proc. 2018 Int. Conf. Current Trends Toward Converging Technol. (ICCTCT), Coimbatore, India, 2018, doi: 10.1109/ICCTCT.2018.8551008.

[9] M. Ghafoor et al., “A Review of Deep Learning Techniques for Lung Cancer Screening and Diagnosis Based on CT Images,” Diagnostics, vol. 13, no. 16, Art. no. 2659, 2023, doi: 10.3390/diagnostics13162659.

[10] J. Jian et al., “A robust deep learning algorithm for lung cancer detection from CT images,” Journal of Biomedical Informatics, vol. 145, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666521225000067sciencedirect.com

[11] Q. Da et al., “Deep Machine Learning for Medical Diagnosis, Application to Lung Cancer Detection: A Review,” Appl. Sci., vol. 14, no. 1, p. 15, 2024, doi: 10.3390/applsci14010015.

[12] S. Alfantookh et al., “Deep learning-based approach to diagnose lung cancer using CT-scan images,” Comput. Intell. Neurosci., 2024, doi: 10.1016/j.cineu.2024.100169.

[13] R. Nair and V. Veena, “A comparative study of lung cancer detection using machine learning algorithms,” in Proc. 2019 IEEE Int. Conf. Electr., Comput. Commun. Technol. (ICECCT), Coimbatore, India, 2019, doi: 10.1109/ICECCT.2019.8869001.

[14] H. F. Al-Yasriy, M. S. Al-Husieny, F. Y. Mohsen, E. A. Khalil, and Z. S. Hassan, "Diagnosis of Lung Cancer Based on CT Scans Using CNN," IOP Conference Series: Materials Science and Engineering, vol. 928, 2020.

[15] H. F. Kareem, M. S. A.-Husieny, F. Y. Mohsen, E. A. Khalil, and Z. S. Hassan, “Evaluation of SVM performance in the detection of lung cancer in marked CT scan dataset,” Indonesian Journal of Electrical

Engineering and Computer Science, vol. 21, no. 3, pp. 1731-1738, 2021, doi: 10.11591/ijeecs.v21.i3.pp1731-1738.

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