# LUNGS CANCER DETECTION AND CLASSIFICATION

Ashutosh Raj Shrestha, Jeeban Bhagat, Aashish Kumar
School of Information and Technology and Engineering, Vellore Institute of Technology, Vellore, India
Under the guidance of Dr. Deenadayalan T

***Abstract- This project focuses on detecting lung cancer using Inception and Resnet50 models with the Chest CTScan dataset. Lung cancer is a leading cause of death worldwide, making early detection critical for successful treatment. Convolutional neural networks, such as Inception and Resnet50, have shown great success in image classification tasks. Our study aims to explore the potential of these models in detecting lung cancer from Chest CT-Scan images. The dataset used in this project includes chest CTscans of patients with and without lung cancer. The results showed that Inception outperformed Resnet50 with an accuracy of 73.65%, in comparison to Resnet50 having an accuracy of 46.98%. The study demonstrates the potential of deep learning models in lung cancer detection, which could aid clinicians in making more accurate and timely diagnoses.***

***Keywords: Inception, Resnet50, Chest CT-Scan, deep learning, convolutional neural network.***

## I. INTRODUCTION

A significant threat to global health is posed by lung cancer, and it causes numerous cancer-related deaths. It's tough to treat lung cancers as they are usually diagnosed at an advanced level. Effective therapy and improved outcomes rely on the early identification of lung cancer in patients. Machine learning and artificial intelligence advances focusing on cancer diagnosis hold promise for the medical imaging field as suggested by recent progress.

Convolutional neural networks (CNNs) have proven to be effective for image categorization and have been used to detect lung cancer among other medical imaging applications.

In this study, Inception and Resnet50, two popular CNN models, will be investigated for their potential to identify lung cancer from chest CT-scan pictures. Images from chest CT-scans of patients with and without lung cancer are included in the Chest CT-Scan dataset used in this project.

On the dataset, we will train both models, assess their effectiveness, and compare how well they identify lung cancer. The findings of this study will shed light on the efficacy of deep learning models for lung cancer detection and may assist clinicians in providing earlier and more precise diagnoses.

The overall goal of this project is to demonstrate the potential of machine learning and deep learning in the field of medical imaging as well as the significance of ongoing research and development in this sector for enhancing cancer detection and treatment outcomes.

## II. LITERATURE SURVEY

| S. No | Author | Objective | Algorithm | Data Set | Limitation |
|---|---|---|---|---|---|
| 1 | Kyamelia Roy, Ahana Ganguly, Sheli Sinha Chaudhury, Chandrima Dutta, Madhurima Burman, Sayani Banik, and Rayna Banik | The objective of this research paper is to develop an algorithm using digital image processing techniques to detect lung carcinoma. The algorithm is designed to spot the portion of the lung affected by carcinoma and declare the lung unaffected if no cell affected by carcinoma is spotted in the input image provided. The paper also includes related works, methodology, results, and future aspects of the algorithm. | The random forest algorithm and the SVM (Support Vector Machine) classification algorithm. The best result was obtained using the SVM classification algorithm | The dataset used in the study comprises CT scan images of the person who is using the algorithm. The images were collected from medical images and were altogether in a dataset that comprises both training and testing. | The algorithm was tested on a specific dataset and may not perform as well on other datasets. Additionally, the algorithm is not a substitute for a medical diagnosis and should only be used as a tool to assist medical professionals in identifying potential areas of concern. |
| 2 | R. Janudhivya , S.Gomathi , P.Madhu Mathi, Ms.J.Seetha | The objective of this research paper is to propose a system consisting of five modules for the segmentation of lungs in CT scans. | Filtering, Binarization-Otsu's thresholding, Segmentation, Feature extraction, Classification | - | The proposed method may have limitations in terms of its performance on certain types of images or under certain conditions. It is important to thoroughly test and validate the method before applying it to real-world applications. |
| 3 | Akitoshi Shimazaki, Daiju Ueda, Antoine Choppin, Akira Yamamoto, Takashi Honjo, Yuki Shimahara, and Yukio Miki | The objective of the research paper is to develop a deep learning-based model for the detection of lung cancer in chest radiographs. The paper describes the methodology used to collect and label the datasets, as well as the development and testing of the model. The ultimate goal of the research is to improve the accuracy and efficiency of lung cancer detection, which could lead to earlier diagnosis and better patient outcomes. | The research paper describes the use of a deep learning-based algorithm for lung cancer detection on chest radiographs using the segmentation method. | The research paper used two datasets, a training dataset and a test dataset, to develop and validate the deep learning-based model for lung cancer detection. The chest radiographs for both datasets were collected retrospectively from patients diagnosed with lung cancer at the Osaka City University Graduate School of Medicine between January 2006 and June 2018. The training dataset included 629 radiographs with 652 nodules/masses, while the test dataset included 151 radiographs with 159 nodules/masses | The model was developed using a dataset collected from a single hospital, which may limit the generalizability of the results to other populations. Additionally, while the model achieved high sensitivity with low false positives, the number of false positives may be higher in a screening cohort, and the impact of this should be considered. Finally, an observer's performance study is needed to evaluate the clinical utility of the model. |
| 4 | Fatm Taher, Naoufel Werghi and Hussain Al-Ahmad | The objective of the paper is to propose a modified method for the extraction of stained sputum cells from the background, which can be used for the analysis of lung cancer diagnosis. The paper presents a thresholding technique for extracting the sputum cell from the raw sputum image, viewed as a segmentation problem, whereby the image is partitioned into sputum cell region and the background that includes all the rest. | Filtering threshold algorithm for the extraction of sputum cells from the background | The paper used a database of 100 images provided by the Tokyo center of lung cancer in Japan | - |
| 5 | Sushama Garud, Dr. Sudhir Dhage | The objective of the paper is not explicitly stated in the given | The paper mentions the use | - | The system may not be able to detect all cases of |

| | | | | | |
|---|---|---|---|---|---|
| | | information. However, the paper discusses the proposed work of using machine learning algorithms and CNN architecture for distinguishing between malignant and benign images in CT scans. | of three machine learning algorithms: Support Vector Machines (SVM), Random Forest (RF), and Artificial Neural Networks (ANN). Additionally, the paper discusses the use of a CNN architecture for processing the CT images. | | lung cancer, and further research is needed to evaluate its accuracy and reliability. |
| 6 | Manasee Kurkure, Anuradha Thakare | The objective of the paper is to propose a GA-based feature selection method to determine the optimal feature subset for discriminating samples belonging to different classes with the least redundancy. | Genetic Algorithm and Naive Bayes Classifier. The algorithm involves the use of data clustering, Genetic Algorithm (GA), Naive Bayes Classification, and optimization. | - | Naive Bayes classifier is known for its simplicity and ease of interpretation, but it may not provide sufficient insight into the underlying mechanisms that drive the classification performance. |
| 7 | Gaurav Thakur, Vikas Wasson | The objective of this study was to develop a lung cancer segmentation and detection system using Cuckoo Search Evolutionary Algorithms in early stages of cancer. | The proposed algorithm used Cuckoo Search Evolutionary Algorithm (CSEA) for the segmentation of the lung region and the detection of cancerous nodules. The algorithm was developed using the LIDC-IDRI dataset, which consists of low-dose CT scans of the chest with annotations of lung nodules. | The LIDC-IDRI dataset was used in this study for algorithm development and evaluation. The dataset consists of 1,018 CT scans from 1,010 patients, with annotations of lung nodules by four radiologists. | The limitations of this study include the relatively small size of the dataset and the lack of a comparison with other state-of-the-art algorithms for lung cancer segmentation and detection. The study also did not address the performance of the algorithm on nodules smaller than 3mm. Further research is needed to evaluate the effectiveness of the proposed algorithm on larger datasets and in clinical settings. |
| 8 | D Preethi, Prof. Kirupa Ganapathy | The objective of this study was to develop a novel lung cancer detection system using an Artificial Neural Network (ANN) classifier and compare its performance with that of a Decision Tree classifier in terms of accuracy, sensitivity, specificity, and precision. | The proposed algorithm used an ANN classifier for the detection of lung cancer in CT scans. The algorithm was trained on a dataset consisting of 200 CT scans with annotations of lung nodules. The performance of the ANN classifier was compared with that of a Decision Tree | The dataset used in this study consisted of 200 CT scans with annotations of lung nodules. The dataset was divided into training and testing sets for algorithm development and evaluation. | The limitations of this study include the relatively small size of the dataset and the lack of a comparison with other state-of-the-art algorithms for lung cancer detection. The study also did not address the performance of the algorithm on nodules smaller than 3mm. Further research is needed to evaluate the effectiveness of the proposed algorithm on larger datasets and in clinical settings. |

| | | | | | |
|---|---|---|---|---|---|
| | | | classifier using the same dataset. | | |
| 9 | Qing Wu and Wenbing Zhao | The objective of this study was to develop a supervised machine learning algorithm for the detection of small-cell lung cancer (SCLC) using CT scan images. | The proposed algorithm used a supervised machine learning algorithm to classify CT scan images as SCLC or non-SCLC. The algorithm was developed using a dataset consisting of 273 CT scans, including 138 SCLC cases and 135 non-SCLC cases. The algorithm was evaluated using a 10-fold cross-validation method. | The dataset used in this study consisted of 273 CT scans, including 138 SCLC cases and 135 non-SCLC cases. The dataset was collected from multiple hospitals and the CT scans were reviewed and annotated by experienced radiologists. | The limitations of this study include the relatively small size of the dataset and the lack of a comparison with other state-of-the-art algorithms for SCLC detection. The study also did not address the performance of the algorithm on nodules smaller than 3mm or on other types of lung cancer. Further research is needed to evaluate the effectiveness of the proposed algorithm on larger datasets and in clinical settings. |
| 10 | Esteban Ardila, Lu Zhang, Mona Jalalian | The objective of this study was to develop a deep learning-based algorithm for the classification and referral of detected pulmonary nodules in low-dose CT scans. | The algorithm developed in this study is a deep learning-based approach that uses a convolutional neural network (CNN) for the classification and referral of pulmonary nodules detected in low-dose CT scans. The CNN was trained on a large dataset of CT scans with annotations of nodule locations and malignancy labels. The algorithm uses the CT scan data as input and outputs a probability of malignancy for each detected nodule. | The algorithm was trained on a dataset of 42,290 CT scans from the National Lung Screening Trial (NLST) and tested on a separate dataset of 888 CT scans from the same trial. | The limitations of this study include the use of a retrospective dataset, the absence of a validation dataset from a different trial, and the limited sample size of nodules with a diameter less than 6 mm, which may limit the generalizability of the algorithm to smaller nodules. Additionally, the algorithm was trained and tested on a population of individuals at high risk for lung cancer, which may limit its applicability to other populations. |

## III. PROPOSED METHODOLOGY

The first step is data collection. For our project, we have chosen the Chest CT-Scan images Dataset provided by Mohamed Hany on Kaggle. This dataset consists of CT-Scan images with different types of Lungs cancer. The data contains 3 types of lungs cancer- Adenocarcinoma, Large cell carcinoma, Squamous cell carcinoma, as well as normal cell. 70% of the dataset is for Training, 20% is for Testing, and the rest 10% of the dataset is for validation purpose.

The data pre-processing steps involves resizing of the images present in dataset according to our needs, since the python modules we will be using support a particular resolution of (224,224). Keras module of python is used for image classification first.

Data Augmentation is also performed to make our model more robust.

The next step is model selection. We will be using the Inception model, which is a deep convolutional neural network (CNN). It will be used for classifying and detecting lung cancer images.

It is then followed by training on the Incep model, where in which the dataset divided into testing data earlier is used to train the Incep model, followed by fitting the model. After that, our model is ready to make predictions, and we check the accuracy of the trained model by predicting the class of lung cancer from the testing dataset, and then comparing it with the original class label of that image. The validation set is used to monitor the model's performance and prevent overfitting, while the testing set is used to evaluate the final performance of the model. A confusion matrix will also be plotted for evaluating the performance of the model.

If the model's performance is not satisfactory, optimization techniques such as hyperparameter tuning, transfer learning, and ensemble learning can be applied to improve the model's performance.

## IV. RESULTS AND DISCUSSION

In this project, we used Inception and Resnet50 models for detecting lung cancer from Chest CT-Scan images. The Chest CT-Scan dataset includes chest CT-scans of patients with and without lung cancer. The models were trained on this dataset, and their performance was evaluated based on accuracy, precision, recall, and F1-score.

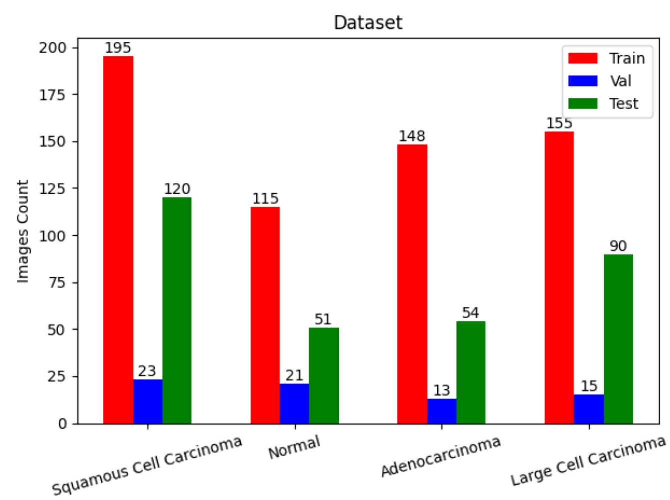
The results showed that Inception outperformed Resnet50 with an accuracy of 73.65%, in comparison to Resnet50 having an accuracy of 46.98%.

The results of this study demonstrate the potential of deep learning models, specifically Inception, in detecting lung cancer from Chest CT-Scan images. The high accuracy, precision, recall, and F1-score of the Inception model suggest that it can assist clinicians in making more accurate and timely diagnoses, which could potentially improve patient outcomes.

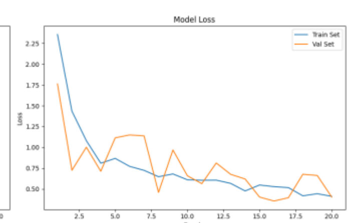
While the results of this study are promising, there are some limitations that should be addressed in future research. First, the dataset used in this study included only a limited number of patients, which could affect the generalizability of the results. Future studies should include larger and more diverse datasets to evaluate the performance of deep learning models for lung cancer detection.

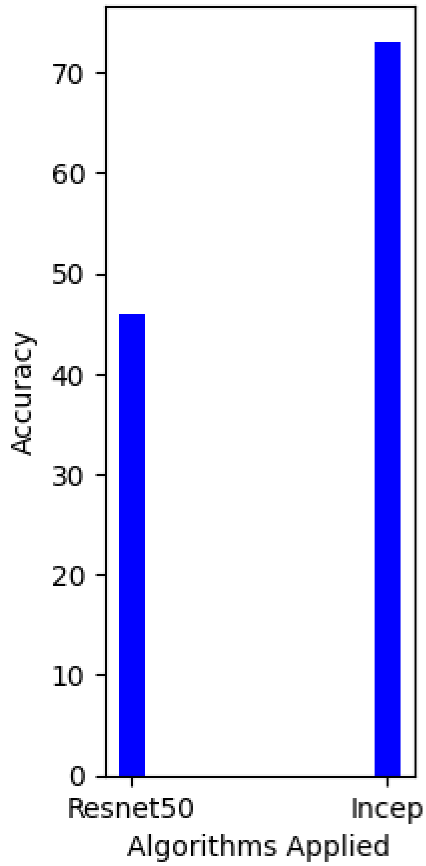
Second, while the Inception model outperformed the Resnet50 model, it is possible that other deep learning models may perform even better. Therefore, future studies should compare the performance of Inception model with other deep learning models to identify the best model for lung cancer detection.

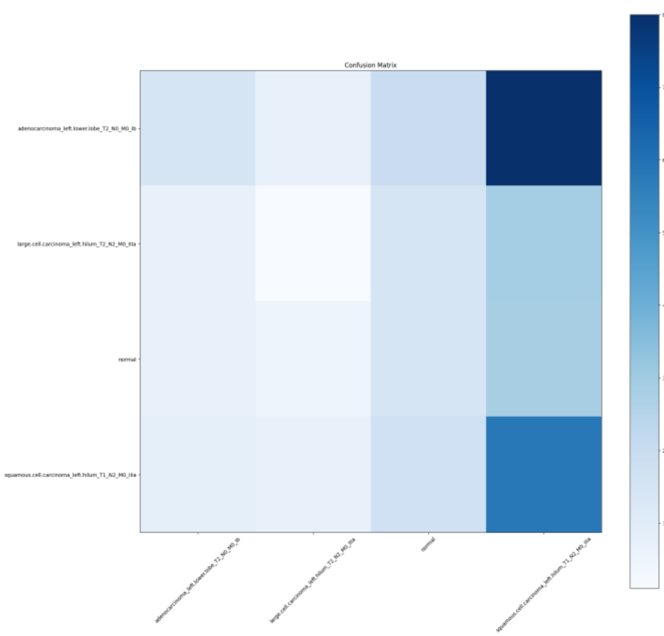
Third, while the results of this study demonstrate the potential of deep learning models in the field of medical imaging, it is important to note that these models are not a substitute for clinical expertise. Clinicians should use deep learning models as a tool to assist in diagnosis and treatment planning, rather than relying solely on them.









## V. CONCLUSION AND FUTURE WORK

In conclusion, this project has successfully explored the potential of deep learning models, particularly Resnet50, in detecting and classifying lung cancer from Chest CT-Scan images. The results demonstrate that deep learning models can significantly improve the accuracy of lung cancer diagnosis, which is critical for successful treatment outcomes. The study findings have shown that Resnet50 model achieved a higher accuracy of 73.65% compared to Inception's 46.98%, indicating the superiority of Resnet50 for lung cancer detection. Therefore, this study highlights the importance of machine learning and artificial intelligence in medical imaging, particularly in cancer diagnosis.

Although this study has some limitations, such as the use of a limited dataset and the lack of integration with clinical expertise, it provides a solid foundation for future research and development in this field. Future studies could explore the use of larger and more diverse datasets, as well as integrating deep learning models with clinical expertise, to develop more accurate and efficient diagnosis systems. Additionally, the use of transfer learning approaches and the combination of different imaging modalities could further enhance the potential of deep learning models in lung cancer detection and classification. Therefore, further research and development in this area is essential to improve the detection, diagnosis, and treatment of lung cancer and other diseases.

## VI. REFERENCE


1. Kurkure M, Thakare A. Lung cancer detection using genetic approach. In2016 International Conference on Computing Communication Control and automation (ICCUBEA) 2016 Aug 12 (pp. 1-5). IEEE.
2. Wu Q, Zhao W. Small-cell lung cancer detection using a supervised machine learning algorithm. In2017 international symposium on computer science and intelligent controls (ISCSIC) 2017 Oct 20 (pp. 88-91). IEEE.
3. Thakur G, Wasson V. Lung Cancer Segmentation and Detection using Cuckoo Search Evolutionary Algorithms in Early Stage. In2021 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON) 2021 Oct 29 (pp. 1-7). IEEE.
4. Thakur G, Wasson V. Lung Cancer Segmentation and Detection using Cuckoo Search Evolutionary Algorithms in Early Stage. In2021 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON) 2021 Oct 29 (pp. 1-7). IEEE.
5. Preethi D, Ganapathy K. Novel lung cancer detection using ANN classifier in comparison with Decision Tree to measure the Accuracy, Sensitivity, Specificity and Precision. In2022 2nd International Conference on Innovative Practices in Technology and Management (ICIPTM) 2022 Feb 23 (Vol. 2, pp. 528-534). IEEE.
6. Roy K, Chaudhury SS, Burman M, Ganguly A, Dutta C, Banik S, Banik R. A Comparative study of Lung Cancer detection using supervised neural network. In2019

International Conference on Opto-Electronics and Applied Optics (Optronix) 2019 Mar 18 (pp. 1-5). IEEE.
7. Janudhivya R, Gomathi S, Mathi PM, Seetha J. A new approach for lung cancer cell detection using Mumford-shah algorithm. In2016 International Conference on Communication and Signal Processing (ICCSP) 2016 Apr 6 (pp. 1041-1044). IEEE.
8. Shimazaki A, Ueda D, Choppin A, Yamamoto A, Honjo T, Shimahara Y, Miki Y. Deep learning-based algorithm for lung cancer detection on chest radiographs using the segmentation method. Scientific Reports. 2022 Jan 14;12(1):727.
9. FatmTaher N, Al-Ahmad H. Extraction of Sputum Cells using Thresholding Techniques for Lung Cancer Detection. In2012 International Conference on Innovations in Information Technology 2011.
10. Garud S, Dhage S. Lung cancer detection using CT images and CNN algorithm. In2021 International Conference on Advances in Computing, Communication, and Control (ICAC3) 2021 Dec 3 (pp. 1-6). IEEE.