Cancerous Lung Nodule Detection Using Deep Learning

[1]Ranjitha U N, [2]Deeksha H Kottary, [3]Hasitha Chennubhotla, [4]Rachitha Padela, [5]Bhumika R

Authors affiliation, [1]ranjitha.un@reva.edu.in, [2]deekshahk20@gmail.com, [3]hasitha113@gmail.com, [4]rachithapadela4@gmail.com, [5]bhumikaraj04@gmail.com

Abstract.

A contagious lung tumor is caused by irresistible tissue development in the lungs, this is commonly called Lung Cancer and in biological terms called the Lung Carcinoma. We aim to use Deep Learning - CNN to identify cancerous lung nodules in the CT scan images. The LUNA16 dataset is the primary dataset we've used. A model that can identify the cancerous lung nodules and can distinguish the particular regions in the CT scan images will be our final product. We have used a part of the LUNA16 data set which is a smaller data set which has not been used previously. We are training our model with a smaller dataset because it is not possible for everyone to download a large dataset to train a model. Hence, when we are finding how efficiently a model works if trained with a smaller dataset. So, after training the model, from the experimental results we have achieved an accuracy of 96%.

For English language already ANPR system is available. For kannada language number plate we are proposing this model.

Keywords. Deep CNN, Lung Cancer Detection, Computer Tomography Scan, LUNA16 Dataset.

1. INTRODUCTION

One of the supreme causes of cancer-related deaths across the world is Lung cancer. In the year 2021, there were 130,180 confirmed deaths out of 236,740 fresh lung carcinoma cases. However, early recognition of malignant lung nodules is vital for a favorable prognosis and to reduce mortality percentage. In simple words, lung cancer is the unregulated division of unwanted cells in the lung nodules. Pulmonary nodules are tiny cell growths inside the lungs that can be cancerous (malignant) or noncancerous (benign). Malignant cells are cells that grow out of control, invading neighboring tissues and spreading to other regions of the body via the blood and lymph system. One of the leading causes of cancer in the lungs is smoking. Some of the symptoms of lung cancer are: Facing trouble while breathing, noticing blood when you cough, Pain in the chest, Weakness etc. Studies have shown that if cancer in the lungs is diagnosed in the initial stages, there are

higher chances of it being cured successfully and cancer prognosis is significantly improved. However, determining the possibility of malignancy in early malignant lung nodules is a challenging problem.

MRI, biopsies, or surgeries are not suitable for moving organs such as lungs since the lung is a fragile and delicate organ and intrusive techniques pose a high risk of infection and raise patients' anxiety. Furthermore, it is expensive and takes a lot of time to give us results. CT imaging is the best approach to examine the diseases in the lungs because it offers highly detailed images of a large variety of tissues. CT scan pictures are suitable for analyzing lung cancer because they have higher magnification and can identify calcification. nevertheless, only 68 per cent of lung cancer nodules are successfully diagnosed when only one radiologist analyses the CT scan personally. Moreover, a radiologist's workload increases dramatically while evaluating a CT scan personally, for the presence of a nodule since detection efficiency is affected by nodule characteristics such as size, position, form, structures, and density.

To deal with this problem, our initial step will be to segment CT scans after preprocessing. After masking the CT scan, the subsequent step is to develop a model on the Deep Convolutional Neural Network (DCNN) model to achieve extreme precision. Early studies using deep neural networks for applications in medical images successfully demonstrated improvements in segmentation tasks. A part of the LUNA16 dataset will be used as the input layer in this project. This particular small dataset has never been used for training and we are using this dataset to test how efficiently we can train the model with a smaller dataset. The LUNA16 dataset is a subdivision of the LIDC-IDRI dataset which comprises 1,186 lung nodules labelled in 888 CT images.

2. LITERATURES SURVEY

The paper uses the lung cancer images dataset from the data world and uses classification algorithms like SVM, Logistic Regression, Naive Bayes, and Decision Tree to analyses and implement lung cancer prediction. Radhika P.R and her team achieved an accuracy of 66.7 % on Logistic Regression, 87.87 % on Naive Bayes and 90% on a Decision tree. [1]

This research paper contains a model which is a combination of Convolutional neural networks and ML algorithms like XG Boost, Support Vector Classifier and Random Forest that finds cancer in the lungs using cytopathology images. Basra Jehangir and his team achieved an accuracy of 99.13%. [2]

This paper uses the LIDC-IRDI dataset and uses XGBoost and Random Forest algorithms. Siddharth Bhatia and his team achieved an overall accuracy of 84%. [3]

This paper used the NLST dataset and applied Deep Scanner, a Deep learning-based newly developed algorithm. This model can predict a patient’s cancer status with an accuracy of 78.2%. [4]

In this research, Deep CNN was utilized to detect malignant and noncancerous lung nodules using CT images from the LIDC-IDRI dataset for lung cancer categorization. Deep convolutional neural networks are one of the finest algorithms for lung cancer diagnosis, according to Amjad Khan and his team, who reached a 100% accuracy rate. [5]

This paper worked on their self-made dataset and applied the concepts of VGG16-T a Deep convolutional neural network. In this, they detect pathological lung cancer at an early

stage. Shenzhen Pang and his team achieved an overall accuracy of 84%. [6]

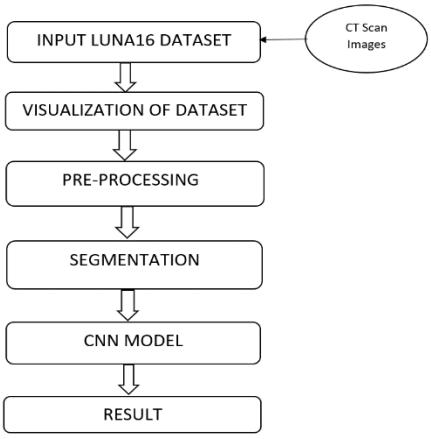
3. CHALLENGES

The dataset was problematic to download since the resources are hard to obtain and the hardware for processing the data is inadequate since the machinery for processing the data was insufficient due to the low quantity of storage available.

When the data is encoded to a 2D grayscale picture, it becomes difficult to train it with the TensorFlow and Keras frameworks since such forms are not supported during model development.

The dataset we used was in. mdh format, which is different from typical image processing formats like.jpg and so on. This layout stores all or most of the patient's information as well as a 3D CT-Scan image of the patient, however when converted to a 2D NumPy image results in the shape of (n, m, m, 1), which is not supported by transfer learning models.

4. METHODOLOGY



4.1. Dataset

We used a small part of the LUNA16 dataset which is the subset of the LIDC-IDRI dataset. It contains Dicom formatted CT scan of the lungs. It has annotated around 888 CT scans from 1186 lung nodule images.

4.2. Visualization

Visualization of the dataset is an important part of training, as it gives a better

understanding of the dataset. CT scan images are difficult to view on a standard computer or in any window browser. We utilize the Simple ITK library to solve this problem.
So here we firstly read the images and retrieve their voxel coordinates, resolution and their origins which is further used in finding the region of interest. Dicom images have patient information and doctor information in them. We use cv2 to resize the image dimensions to (224,224), which helps in further processing.

4.3. Pre-Processing: Segmentation

Pre- Processing any data in an ML model is a very crucial step. Pre-Processing of the image is done in order to enrich some features in the image and modify it to analyze and train it better. The CT scans are firstly split to training and testing dataset with a ratio of 80:20. Now these CT scans are segmented which is one of the Pre-Processing methods.
To find the region of interest in the lung CT scan image, we must perform segmentation on it which is an important step. Therefore, to make the feature extracting process more functional, segmentation of the Lung CT scan is done. To single out the lung region, connected component analysis can be used to segment out the largest air pocket in the lung. We normalize the CT scans for better homogeneity.
We are using the C3D based architecture with added adjustments to build the feature extractor, which will be used to predict the malignant lung nodules. The BRISK method is a feature point identification and description technique with scale and Preserving Parallelism-Rotations invariance. It generates the binary feature descriptor by constructing the feature descriptor of the local picture using the grey scale relationship of random point pairs in the local image's neighborhood. When compared to the old approach, BRISK's matching performance is quicker and its storage memory becomes less, but its robustness is decreased. Initially, we pre-train the model using the CT scan of some patients and then using the other CT scan images of other patients which was pre-processed we further train the model.

4.4. Model Implementation

We are using the Deep Learning Algorithm called the DCNN with Fully Connected Layers and Max Pooling. This is utilized for feature extraction and categorization of lung cancer cases that are positive or negative.
The feature extraction is done in many processes, each of which has three cascade layers namely the Convolutional Layer, Activation Layer, and Max-Pooling Layer.

1. Layer 1: Image is sent or fed to the network in this layer, and it is called the Input Layer.
2. Layer 2: The most significant aspect of DCNN is the convolutional layer. Its duty is to extract features from the data it receives. Convolution must be performed using a succession of kernels in these levels.
3. Layer 3: This layer adds nonlinearity to the algorithm, making it easier to understand complicated data and is called the Activation Layer.
4. Layer 4: The non-overlapping section of the original region is max-pooled by using a max filter and is called the Max-Pooling Layer.

We are also applying the Adam optimization. Adam is one of the best optimizers and can handle sparse gradients on noisy problems. We use it for the classification of the positive and negative cases of lung cancer. By applying this model, we have achieved an accuracy of 96%.

5. RESULTS

In this project, we used the Deep Convolutional Model 5 Neural Network. With this model, we got the following Accuracy, Validation Accuracy, Loss, Validation Loss.

The following is the Classification report we achieved:

	PRECISION	RECALL	F1-SCORE
0	0.99	0.96	0.98
1	0.84	0.98	0.90
ACCURACY			0.96
MACRO AVG	0.92	0.97	0.94
WEIGHTED AVG	0.97	0.96	0.96

Here '1' represents the positive cases and '0' the negative cases. Precision is the performance of the model in predicting true positives to false positives and true negatives to false negatives. Recall shows the performance of identifying true positives out of the total true positive cases and identifying true negatives out of the total true negative's cases. If we take the harmonic mean of precision and recall, we get the F-score.
We have used a part of the LUNA - 16 data set which is a smaller data set which has not been used previously. We are then checking for accuracy. After training the model, we have achieved precision on negative cases up to 99%, precision on positive cases up to 84% and total accuracy of 96%.

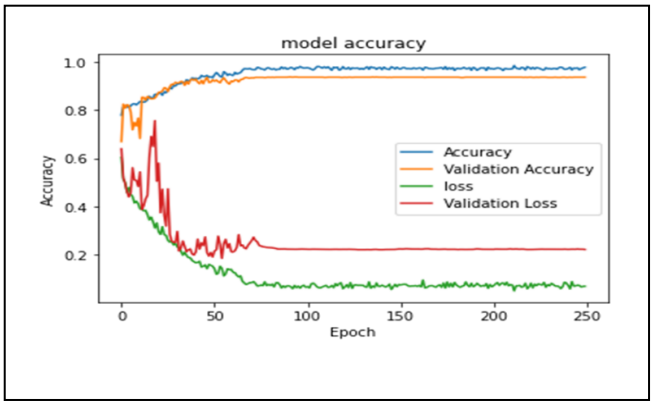
6. FUTURE WORK

In order to increase the accuracy, we can implement a combination of the Vgg-16 algorithm and Deep convolutional neural networks. This might increase the precision and the recall metrics as well as we are using a smaller dataset. Using a larger dataset would help in training the model better and make best use of the model. We would also like to work with different types of images and datasets and observe how the model behaves.

7. CONCLUSION

The project's ultimate goal was to test and improve the accuracy of lung nodule diagnosis while trained with a smaller dataset and prediction of positive cases using CT scan of the lungs. The promising results from the previous study encourage us to test the CNN-based lung cancer detection model on larger datasets. We wanted to see how well this model works on a smaller dataset. We have come to a conclusion that DCNN works better on larger datasets than on smaller datasets. Nevertheless, DCNN proves to be very accurate

and makes it easier for the radiologists to detect malignant nodules with utmost accuracy. We have implemented the DCNN Model which is giving us an accuracy of 96%.



8. REFERENCES

[1] Sreekumar, Amrit, et al. "Malignant Lung Nodule Detection using Deep Learning." 2020 International Conference on Communication and Signal Processing (ICCSP). IEEE, 2020.

[2] Radhika,P.R., Rakhi AS Nair, and G Veena "A comparative study of lung cancer detection using machine learning Algorithms." 2019 IEEE International conference on Electrical,Computer and Communication Technologies (ICECCT). IEEE,2019.

[3] Heeneman, Thomas, and Mark Hoogendoorn. "Lung nodule detection by using deep learning." University of Amsterdam, Research Paper (2018).

[4] Jehangir,Basra, Soumya Ranjan Nayak and Sourav Shandilya. "Lung Cancer detection using machine learning Models" 2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence), IEEE, 2022.

[5] Bhatia, Siddharth, Yash Sinha and Lavika Goel. "Lung Cancer detection: a deep learning approach." Soft Computing for Problem Solving. Springer, Singapore, 2019. 699-705.

[6] Khan, Amjad: "Identification of Lung Cancer Using Convolutional Neural Networks Based Classification" Turkish Journal of Computer and Mathematics Education (TURCOMAT) 12.10 (2021): 192-203.

[7] Causey Jason L., et al "Lung Cancer screening with low dose CT scans using deep learning approach. "airXiv preprint airXiv :1906.00240(2019).

[8] Vas, Moffy, and Amita Dessai. "Lung cancer detection system using lung CT image processing." 2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA). IEEE, 2017.

[9] A. Soni and A. P. Singh, "Automatic Pulmonary Cancer Detection using Prewitt & Morphological Dilation," 2nd International Conference on Data, Engineering and Applications (IDEA), 2020, pp. 1-6, doi: 10.1109/IDEA49133.2020.9170680.

[10] M. Šarić, M. Russo, M. Stella, and M. Sikora, "CNN-based Method for Lung Cancer Detection in Whole Slide Histopathology Images," 2019 4th International Conference on Smart and Sustainable Technologies (SpliTech), 2019, pp. 1-4, doi: 10.23919/SpliTech.2019.878304

[11] N. Nawreen, U. Hany and T. Islam, "Lung Cancer Detection and Classification using CT Scan Image Processing," 2021 International Conference on Automation, Control and Mechatronics for Industry 4.0 (ACMI), 2021, pp. 1-6, doi: 10.1109/ACMI53878.2021.9528297.

[12] Y. Balagurunathan et al., "Lung Nodule Malignancy Prediction in Sequential CT Scans: Summary of ISBI 2018 Challenge," in IEEE Transactions on Medical Imaging, vol. 40, no. 12, pp. 3748-3761, Dec. 2021, doi: 10.1109/TMI.2021.3097665.

[13] S. A. D. L. V. Senarathna, S. P. Y. A. A. Piyumal, R. Hirshan and W. G. C. W. Kumara, "Lung Cancer Detection and Prediction of Cancer Stages Using Image Processing," 2021 3rd International Conference on Electrical, Control and Instrumentation Engineering (ICECIE), 2021, pp. 1-9, doi: 10.1109/ICECIE52348.2021.9664658.

[14] N. S. Nadkarni and S. Borkar, "Detection of Lung Cancer in CT Images using Image Processing," 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), 2019, pp. 863-866, doi: 10.1109/ICOEI.2019.8862577.

[15] R. D. Karthikeyan, R. G, V. V, G. B. C and K. M, "A Review of Lung Cancer Detection using Image Processing," 2021 Smart Technologies, Communication and Robotics (STCR), 2021, pp. 1-4, doi: 10.1109/STCR51658.2021.9588835.

[16] Ahmed, S. T. (2017, June). A study on multi objective optimal clustering techniques for medical datasets. In *2017 international conference on intelligent computing and control systems (ICICCS)* (pp. 174-177). IEEE..

[17] S. Mukherjee and S. U. Bohra, "Lung Cancer Disease Diagnosis Using Machine Learning Approach," 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), 2020, pp. 207-211, doi: 10.1109/ICISS49785.2020.9315909.