**ADVANCING EARLY DETECTION: A CNN-BASED APPROACH FOR LUNG CANCER DETECTION IN CT SCAN IMAGES**

# **A PROJECT REPORT**

### **Submitted by**

RUTHRANAYAKI J (21ADR041)

HARIARASU R (21ADR014)

NAVEENAN M (21ADR031)

### 

### 

***for***

# **20ADC33 DEEP LEARNING**

# **DEPARTMENT OF ARTIFICIAL INTELLIGENCE**

**KONGU ENGINEERING COLLEGE**

# **(Autonomous)**

**PERUNDURAI, ERODE – 638 060**

# **NOVEMBER 2023**

**Abstract: As a widespread hazard to world health, lung cancer requires effective early detection techniques. This work focuses on deep learning—more specifically, Convolutional Neural Networks(CNN)—to diagnose lung cancer from CT scan images. Even with CT scans' superior diagnostic capabilities, malignancy detection is still difficult, requiring computer-aided diagnostics. The study methodically assesses several deep learning techniques, emphasizing both their advantages and disadvantages, and ends with the suggestion of an ideal model. To improve performance and prevent overfitting, our model combines a pre-trained ResNet CNN with a 60% dropout layer and a sparsified architecture. Most notably, PyTorch, a flexible deep learning framework, is integrated with ease, promoting experimentation and improving model flexibility. Synthetic Minority Over-sampling Technique (SMOTE) is utilized in conjunction with data augmentation strategies to rectify imbalances in the dataset, guaranteeing the development of a strong training set. Our model's better classification accuracy is demonstrated by rigorous validation utilizing the IQ-OTH/NCCD - Lung Cancer Dataset versus many pre-trained CNNs, such as AlexNet and ResNet50. In addition to improving lung cancer diagnosis, this research highlights PyTorch's critical role in model implementation. SMOTE and other data augmentation together significantly improve our deep learning model's robustness and dependability for precise lung cancer identification in CT scan pictures.**

**Keywords: Lung cancer, deep learning, CNNs, CT scans, computer-aided diagnostics, ResNet, PyTorch, SMOTE, data augmentation, classification accuracy, dataset imbalances, early detection**

**I INTRODUCTION**

Lung cancer is a significant worldwide health concern that demands progress in early detection techniques to enhance patient outcomes. Given its potentially fatal consequences, an accurate diagnosis must be made as soon as possible. The nexus of deep learning and medical imaging has been a game-changer in recent years, especially when it comes to lung cancer diagnosis using CT scan pictures.

The paper discusses the necessity of using deep learning (DL) approaches to enhance diagnostic speed and accuracy. Although CT scans are very useful for providing precise imaging of lung structures, the process of identifying malignant tumors can be difficult to diagnose. Conventional diagnostic techniques frequently fail to identify tiny abnormalities in these photos that point to cancer.

Our study presents a unique DL-based model with the goal of adding to the changing era of diagnosis of lung cancer.Utilizing the strength of CNN, a deep learning algorithms well-known for their ability to interpret pictures, our method combines proven architectures, like ResNet, that have been trained to extract complex information from CT scan images. Furthermore, the solution makes use of the dynamic deep learning library PyTorch framework, which is flexible and efficient.

Our suggested model includes a sparsified architecture with dropout layers to enhance model performance and eliminate overfitting issues. The IQ-OTH/NCCD - Lung Cancer Dataset is a pre-processed CT scan image dataset that is used in the study's rigorous validation of the model. By comparing the results with those of other CNNs that have already been trained, such as AlexNet and ResNet50, our research aims to provide insights into the efficacy and robustness of the proposed deep learning model for the accurate lung cancer detection.

In alignment with the broader trends in medical imaging research, this paper contributes to the ongoing pursuit of precision in lung cancer detection. By combining state-of-the-art DL techniques with advancements in deep learning frameworks, our study aspires to propel the field forward, offering valuable insights for clinicians and researchers alike in the critical domain of early lung cancer diagnosis.

**II LITERATURE REVIEW**

Researchers have made notable progress in the field of lung cancer detection through the analysis of CT images, employing diverse methodologies. Gindi, Al Attiatalla, and Sami [1] developed a model that integrated geometrical, statistical, and gray level characteristics for lung cancer detection. Suzuki et al. [2] focused on radiologically classifying small adenocarcinomas of the lung, establishing correlations between radiological findings and pathological outcomes. Xiuhua, Tao, and Zhigang [3] concentrated on predicting malignant pulmonary nodules based on texture features of CT images. Aggarwal, Furqan, and Kalra [4] implemented feature extraction and LDA-based classification of lung nodules in chest CT scan images with notable results. Jin, Zhang, and Jin [5] utilized a Convolutional Neural Network (CNN) in a CAD system, achieving an accuracy of 84.6%. Sangamithraa and Govindaraju [6] explored K-means clustering for segmentation and back propagation for classification, achieving a commendable accuracy of 90.7%. Roy, Sirohi, and Patle [7] employed a fuzzy inference system and an active contour model for lung cancer nodule detection, showcasing proficiency in utilizing gray transformation for image contrast enhancement. Ignatious and Joseph [8] introduced a system based on watershed segmentation, achieving an accuracy of 90.1%. Gonzalez and Ponomaryvo [9] proposed a system capable of classifying lung cancer nodules as benign or malignant. Miah and Yousuf [10] focused on lung cancer detection from CT images using image processing and neural network techniques. Khobragade et al. [11] explored automatic detection of major lung diseases using Chest Radiographs and classification by feed-forward artificial neural network. To summarize, this literature survey highlights a range of methodologies, from traditional approaches to advanced techniques like CNNs, showcasing the diverse landscape of lung cancer detection research using CT images. Armato et al. [12] contributed to the pool of resources for lung cancer research through the LIDCIDRI dataset. Mamun et al. [13] proposed a lung cancer prediction model using ensemble learning techniques and conducted a systematic review analysis. Makaju et al. [14] addressed lung cancer detection using CT scan images. Hany [4] provided a chest CT-scan images dataset for lung cancer studies. Ausawalaithong et al. [18] explored automatic lung cancer prediction from chest X-ray images using a deep learning approach. Bhandary et al. [19] presented a deep learning framework for detecting lung abnormalities through the study of chest X-ray and lung CT scan images. Da Silva et al. [20] focused on lung nodule diagnosis based on an evolutionary convolutional neural network. Naqi et al. [21] explored lung nodule detection and classification based on geometric fit in parametric form and deep learning. Kalyani et al. [22] proposed an improved lung cancer prediction system using image processing. Roy, Sirohi, and Patle [23] focused on the classification of lung images and nodule detection using a fuzzy inference system. Ignatious and Joseph [24] developed a computer-aided lung cancer detection system. Wason and Nagarajan [25] explored image processing techniques for analyzing CT scan images towards the early detection of lung cancer. This comprehensive literature survey provides insights into diverse approaches, ranging from traditional methods to advanced deep learning techniques, contributing to ongoing efforts to enhance the accuracy and effectiveness of lung cancer detection using CT images.

**III METHODOLOGY**

The process starts with the collection of an image's data from a publically accessible source, which is then pre-processed. The study uses a variety of deep learning models, such as the CNN approach that is recommended and well-known models like conv2d and ResNet-50. Using a dataset of CT images, these models are trained, tested, and validated using the conventional validation technique. In order to determine the deep learning-powered model for accurately identifying malignant, benign as well as normal instances, the acquired data are then calculated and carefully examined. Figure 1 offers a broad overview of the study.

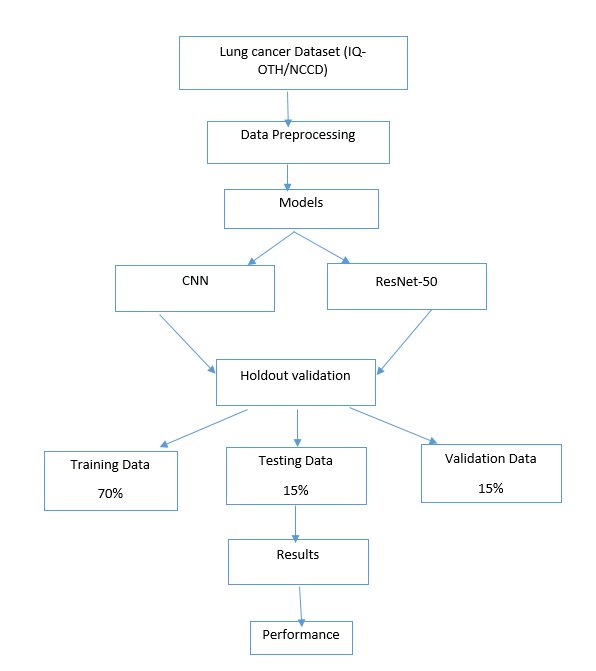
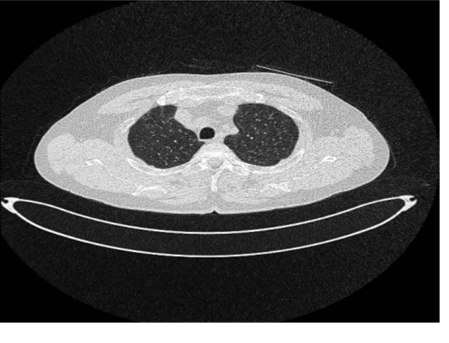
****

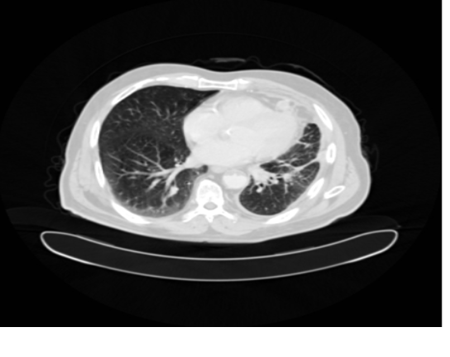
Figure:1 Workflow

**A Data Gathering**

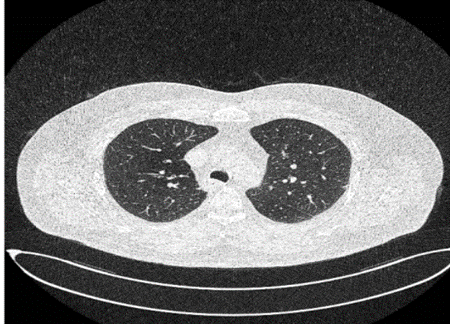
The lung cancer data consisting of computed Tomography (CT) scan images was sourced from the publicly available Kaggle platform. As per the dataset's origin, meticulous efforts were made to hand-collect images from diverse websites, ensuring the verification of each label. Initially acquired in DICOM format, the CT scans underwent conversion.

The dataset contains a total of 1990 images, representing CT scan slices from 110 cases.

* Malignant: 40 cases
* Benign: 15 cases
* Normal :55 cases

**Figure 2: Normal cases**

**Figure 3: Malignant cases**



**Figure 4: Benign cases**

**B Data Pre-processing**

The dataset preprocessing was a multifaceted procedure designed to elevate image quality, involving key steps like feature extraction encompassing image reading, resizing, noise removal (de-noising), image segmentation, and morphology for edge smoothing. These crucial procedures were instrumental in optimizing input data for deep learning models, specifically tailored for image detection tasks. Additionally, the pre-processing pipeline integrated advanced techniques such as SMOTE to address class imbalance and an image generator for augmenting the dataset, enhancing training data robustness. This comprehensive pre-processing laid the groundwork for a more effective and resilient deep learning model. Simultaneously, addressing missing values emerged as a pivotal facet of data pre-processing. This involved managing entries marked as absent, empty, NaN (Not a Number), or another placeholder. Various factors, including measurement errors, data input issues, or the lack of data for specific observations, could contribute

to missing values. Effectively handling these instances was crucial for preserving dataset quality and integrity. Common approaches for addressing missing values included elimination, where rows or columns containing missing values were removed, albeit at the risk of information loss. Alternatively, imputation techniques were employed, replacing missing data with computed or estimated values. Mean, median, or mode imputation involved substituting the respective statistical measures of non-missing values in the same column for any missing values. For more intricate datasets, advanced imputation techniques like regression, k-nearest neighbors, and machine learning-based imputation provided sophisticated alternatives. These combined efforts in pre-processing contributed to a more robust and reliable foundation for subsequent analyses and model training. Handling Missing Values:

One of the most important steps in data preprocessing is handling missing values, which entails dealing with and managing data items that are either unavailable or recorded as empty, NaN (Not a Number), or another placeholder. A number of factors, including measurement errors, problems with data input, or the simple fact that data was not gathered for a specific observation, might result in missing values. Ensuring the quality and integrity of the dataset requires adequate handling of missing values. Typical methods for dealing with missing values consist of:

Removal: In this technique, the dataset's missing values are eliminated from any rows or columns. Although it makes the dataset simpler, information may be lost as a result.

Imputation: Imputation is the process of substituting computed or guessed values for missing data.

Mean, Median, or Mode Imputation: In the same column, substitute the non-missing values' mean, median or mode for any missing values.

Advanced Imputation: Regression imputation, k-nearest neighbor’s imputation, and machine learning-based imputation are a few examples of advanced imputation techniques that can be used for more complicated datasets.

**C Validation Process**

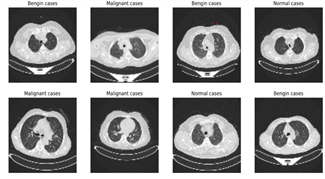
Ensuring the effectiveness of large image datasets hinges on the careful selection of a suitable validation process. In our methodology, we adopted a validation procedure, dedicating 70% for training, 15% for testing, and 15% for validation—a proven and widely accepted method. Across all model, we standardized the epoch’s value to 20 and the batch size to 15. Furthermore, a random seed value of 100 was applied to guarantee result reproducibility, maintaining consistency across multiple iterations.

**D Proposed CNN Architecture**

The envisioned CNN architecture processed a 64x64 input image through an initial convolution layer with 16 filters, producing 62x62 feature to discern fundamental features. Following this, a max-pooling layer with 31x31 feature maps was employed to reduce the spatial data size by half. The process continued with a second convolution layer comprising 32 filters and 29x29 feature maps, succeeded by another max-pooling layer with 14x14 feature maps. A subsequent stage introduced additional convolution and pooling layers, utilizing 64 filters and 10x10 feature maps for the convolution layer and 5x5 feature maps for the pooling layer. The ultimate output underwent flattening and transitioned to a 260-dimensional connected dense layer, followed by a softmax activation function suitable for multiple classifications. All layers, excluding the final one, utilized a ReLU activation function without dropout. The model underwent training, validation, and testing with a learning rate of 0.01, 20 epoch, and a batch size 15, employing the Adam optimizer and cross-entropy as the loss function and callback function is used to define what happens before, during, or at the end of a training epoch

**IV RESULTS**

A comparative analysis of two models, namely CNN and ResNet-50, was conducted on the lung cancer CT scan image dataset. The CNN model demonstrated superior performance, attaining a testing accuracy of 92%, a testing AUC of 98.21%, testing recall of 91.72%, and a testing loss of 0.328. Comprehensive results for training, validation, and testing outputs are delineated in Table I, Table II respectively. The collective findings unequivocally position CNN as the preferred model for detecting cancer in CT scan images.

Figure 5: Types of cases

| **Models** | **Training Accuracy** | **Validation Accuracy** | **Testing Accuracy** |
| --- | --- | --- | --- |
| CNN | 99.80% | 91.11% | 92.00% |
| ResNet | 99.56% | 84.20% | 84.12% |

Table I: Training, Validation and testing accuracy results of different model to detect lung cancer

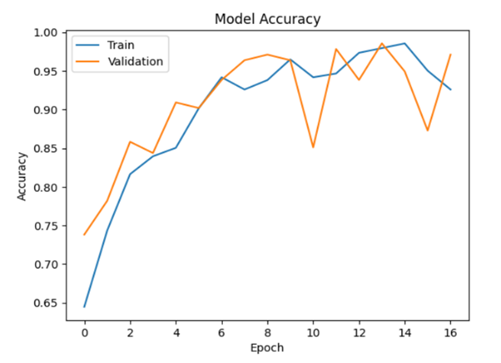


Figure 6: Accuracy

| **Models** | **Training Accuracy** | **Validation Accuracy** | **Testing Accuracy** |
| --- | --- | --- | --- |
| CNN | 0.003% | 0.354% | 0.326% |
| ResNet | 0.043% | 0.596% | 0.596% |

Table II: Training, Validation and testing loss results of different model to detect lung cancer

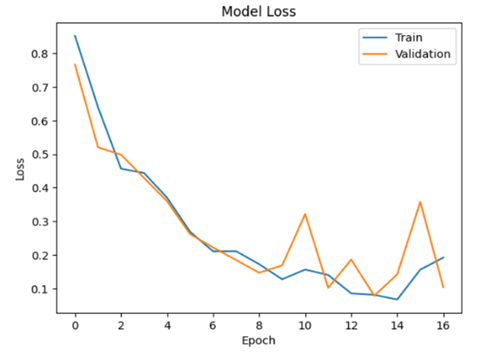


Figure 7: Loss

V CONCLUSION

The global prevalence of lung cancer, coupled with its high mortality rate, underscores the urgent need for early detection to enhance patient outcomes. While lung cancer remains challenging to prevent, prompt diagnosis significantly improves the chances of extended survival. Particularly in developed regions like North America, lung cancer stands as a leading cause of cancer-related deaths. The elusive nature of early-stage detection contributes to the severity of the disease upon discovery. Despite notable advancements, achieving reliable early diagnosis remains a persistent challenge. In this study, we proposed a CNN-based deep learning model for the early detection of lung cancer using CT scan images. Early identification holds the potential for more effective treatment and care, offering the possibility of curing cancer in its initial stages. Beyond the CNN model, we explored alternative models such as ResNet50 and CONV2D. Our comprehensive analysis revealed that the CNN model outperformed others, achieving an accuracy of 92%, AUC of 98.21%, recall of 91.72%, and a loss of 0.328. Moving forward, the integration of additional datasets and the exploration of diverse machine learning and deep learning models could further enhance the early diagnosis of lung cancer.

VI REFERENCES

1) Gindi,A. M., Al Attiatalla, T. A., & Sami, M.M. (2014) “A Comparative Study for Comparing Two Feature Extraction Methods and Two Classifiers in Classification of Earlystage Lung Cancer Diagnosis of chest x-ray images.” Journal of American Science, 10(6): 13-22.

2) Suzuki, K., Kusumoto, M., Watanabe, S. I., Tsuchiya, R., & Asamura, H. (2006) “Radiologic classification of small adenocarcinoma of the lung: radiologic-pathologic correlation and its prognostic impact,” The Annals of Thoracic Surgery. 81(2): 413-419.

3) Xiuhua,G., Tao, S., & Zhigang, L.(2011) “Prediction Models for Malignant Pulmonary Nodules Based-on Texture Features of CT Image.” In Theory and Applications of CT Imaging and Analysis. DOI: 10.5772/14766.

4) Aggarwal, T., Furqan, A., & Kalra, K. (2015) “Feature extraction and LDA based classification of lung nodules in chest CT scan images.” 2015 International Conference On Advances In Computing, Communications And Informatics (ICACCI), DOI: 10.1109/ICACCI.2015.7275773.

5) Jin, X., Zhang, Y., & Jin, Q. (2016) “Pulmonary Nodule Detection Based on CT Images Using Convolution Neural Network.” 2016 9Th International Symposium On Computational Intelligence And Design (ISCID). DOI: 10.1109/ISCID.2016.1053.

6) Sangamithraa, P., & Govindaraju, S. (2016) “Lung tumour detection and classification using EK-Mean clustering.” 2016 International Conference On Wireless Communications, Signal Processing And Networking (Wispnet). DOI: 10.1109/WiSPNET.2016.7566533.

7) Roy, T., Sirohi, N., & Patle, A. (2015) “Classification of lung image and nodule detection using fuzzy inference system.” International Conference On Computing, Communication & Automation. DOI: 10.1109/CCAA.2015.7148560.

8) Ignatious, S., & Joseph, R. (2015) “Computer aided lung cancer detection system.” 2015 Global Conference On Communication Technologies (GCCT), DOI: 10.1109/GCCT.2015.7342723.

9) Rendon-Gonzalez, E., & Ponomaryov, V. (2016) “Automatic Lung nodule segmentation and classification in CT images based on SVM.” 2016 9Th International Kharkiv Symposium On Physics And Engineering Of Microwaves, Millimeter And Submillimeter Waves (MSMW). DOI: 10.1109/MSMW.2016.7537995.

10) Miah, M.B.A., & Yousuf, M.A. (2015) “Detection of lung cancer from CT image using image processing and neural network.” 2015 International Conference on Electrical Engineering and Information Communication Technology (ICEEICT):1-6.

11) Khobragade, S., Tiwari, A., Patil, C., & Narke, V. (2016) “Automatic detection of major lung diseases using Chest Radiographs and classification by feed-forward artificial neural network.” IEEE International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES):1-5.

12) Armato, I., Samuel McLennan, G., McNitt-Gray, F. R., Michael, Charles, Reeves, Anthony P., … Clarke, Laurenc, (2015) “Data From LIDC-IDRI. The Cancer Imaging” Archive.http://doi.org/10.7937/K9/TCIA.2015.LO9QL9SX.

13) M. Mamun, A. Farjana, M. Al Mamun and M. S. Ahammed, "Lung cancer prediction model using ensemble learning techniques and a systematic review analysis," 2022 IEEE World AI IoT Congress (AIIoT), 2022, pp.187193,doi:10.1109/AIIoT54504.2022.9817326.

14) Makaju, Suren, P. W. C. Prasad, Abeer Alsadoon, A. K. Singh, and A. Elchouemi. "Lung cancer detection using CT scan images." Procedia Computer Science 125 (2018): 107-114.

15) A. Farag, J. Graham, A. Farag, Robust segmentation of lung tissue in chest CT scanning, ınternational conference on ımage processing (2010)

16) M. Anthimopoulos, S. Christodoulidis, L. Ebner, A. Christe, s Mougiakakou, lungpattern classification for ınterstitial lung diseases using a deep convolution neural network. IEEE Tran. Med. Imag. **35**, 1207–1216 (2016)

17) Disha, Gagandeep, Identifying lung cancer using ımage processing techniques, ICCTAI, (2011)

18) Ausawalaithong, W., Thirach, A., Marukatat, S., & Wilaiprasitporn, T. (2018). Automatic Lung Cancer Prediction from Chest X-ray Images Using the Deep Learning Approach. 2018 11th Biomedical Engineering International Conference (BMEiCON).

19) Abhir Bhandary , G. Ananth Prabhu , V. Rajinikanth , K. Palani Thanaraj , Suresh Chandra Satapathy , David E. Robbins , Charles Shasky , YuDong Zhang , Joao Manuel R.S. Tavares , N. Sri Madhava Raja , DeepLearning Framework to Detect Lung ˜ Abnormality – A study with Chest X-Ray and Lung CT Scan Images, Pattern Recognition Letters(2019),doi:https://doi.org/10.1016/j.patrec.2019.11.01.

20) Da Silva, G.L.F.; da Silva Neto, O.P.; Silva, A.C.; de Paiva, A.C.; Gattass, M. Lung nodules diagnosis based on evolutionary convolutional neural network. Multimed. Tools Appl. 2017, 76, 19039–19055.

21) Naqi, S.M.; Sharif, M.; Jaffar, A. Lung nodule detection and classification based on geometric fit in parametric form and deep learning. Neural Comput. Appl. 2018, 3456789.

22) D Kalyani, C Raghavendra, K Rajendra Prasad, “An Improved Lung Cancer Prediction System using Image Processing” , International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-4, November 2019.

23) Roy, T., Sirohi, N., & Patle, A. “Classification of lung image and nodule detection using fuzzy inference system.” International Conference on Computing, Communication & Automation. DOI: 10.1109/CCAA.2015.7148560.

24) Ignatious, S., & Joseph, R. “Computer aided lung cancer detection system.” Global ConferenceOn Communication Technologies (GCCT),DOI: 10.1109/GCCT.2015.7342723.

25) Jeyaprakash Vasanth Wason & Ayyappan Nagarajan “Image processing techniques for analyzing CT scan images towards the early detection of lung cancer”, DOI: 10.6026/97320630015596.