A Project report on

Lung Cancer/Disease Detection using Machine Learning

A Dissertation submitted to JNTU Hyderabad in partial fulfillment of the academic requirements for the award of the degree.

Bachelor of Technology

in

Computer Science and Engineering

Submitted by

P. THARUN SAI (20H51A0522)
C. GANESH (20H51A0560)
K. SHIVA ABHIGNA (20H51A05H3)

Under the esteemed guidance of

Mr. A. Vivekanand
(Associate
Professor)

Department of Computer Science and Engineering

CMR COLLEGE OF ENGINEERING& TECHNOLOGY

(UGC Autonomous)
*Approved by AICTE *Affiliated to JNTUH *NAAC Accredited with A^+ Grade

KANDLAKOYA, MEDCHAL ROAD, HYDERABAD - 501401.

2020- 2024

CMR COLLEGE OF ENGINEERING & TECHNOLOGY

KANDLAKOYA, MEDCHAL ROAD, HYDERABAD – 501401

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the Major Project Phase I report entitled **" LUNG CANCER/ DISEASE DETECTION USING MACHINE LEARNING"** being submitted by P. Tharun sai (20H51A0522), C. Ganesh (20H51A0560), K. Shiva Abhigna (20H51A05H3) in partial fulfillment for the award of **Bachelor of Technology in Computer Science and Engineering** is a record of bonafide work carried out his/her under my guidance and supervision.

The results embodies in this project report have not been submitted to any other University or Institute for the award of any Degree.

Mr. A. Vivekanand
Associate Professor
Dept. of CSE

Dr. Siva Skandha Sanagala
Associate Professor and HOD
Dept. of CSE

ACKNOWLEDGEMENT

With great pleasure we want to take this opportunity to express my heartfelt gratitude to all the people who helped in making this project work a grand success.

We are grateful to **Mr. A. Vivekanand, Associate Professor** , Department of Computer Science and Engineering for his valuable technical suggestions and guidance during the execution of this project work.

We would like to thank **Dr. Siva Skandha Sanagala,** Head of the Department of Computer Science and Engineering, CMR College of Engineering and Technology, who is the major driving forces to complete my project work successfully.

We are very grateful to **Dr. Vijaya Kumar Koppula**, Dean-Academics, CMR College of Engineering and Technology, for his constant support and motivation in carrying out the project work successfully.

We are highly indebted to **Major Dr. V A Narayana,** Principal, CMR College of Engineering and Technology, for giving permission to carry out this project in a successful and fruitful way.

We would like to thank the **Teaching & Non- teaching** staff of Department of Computer Science and Engineering for their co-operation

We express our sincere thanks to **Shri. Ch. Gopal Reddy**, Secretary, CMR Group of Institutions, for his continuous care.

Finally, We extend thanks to our parents who stood behind us at different stages of this Project. We sincerely acknowledge and thank all those who gave support directly and indirectly in completion of this project work.

P. THARUN SAI 20H51A0522
C. GANESH 20H51A0560
K. SHIVA ABHIGNA 20H51A05H3

TABLE OF CONTENTS

List of Tables

ABSTRACT

Lung cancer is one of the leading causes of death worldwide. Early detection of this disease increases the chances of survival. Computer-Aided Detection (CAD) has been used to process CT images of the lung to determine whether an image has traces of cancer. This project presents an image classification method based on the hybrid Convolutional Neural Network (CNN) algorithm and Support Vector Machine (SVM). This algorithm is capable of automatically classifying and analyzing each lung image to check if there is any presence of cancer cells or not. CNN is easier to train and has fewer parameters compared to a fully connected network. we came out with CNN-SVM because it gives good performance compared with other results. This method helps for better merit and its ability to classify lung cancer in CT images accurately.

CHAPTER 1
INTRODUCTION

1. INTRODUCTION

1.1. Introduction

Lung cancer detection refers to the process of identifying the presence of lung cancer or the likelihood of its occurrence in an individual's lungs. This detection can be achieved through various medical techniques and diagnostic tools and even with machine and deep learning techniques. Lung cancers usually are grouped into two main types called small cell and non-small cell (non-small cell includes adenocarcinoma and squamous cell carcinoma). These types of lung cancer grow differently and are treated differently. Non-small cell lung cancer is more common than small cell lung cancer [1].

Lung cancer is one of the causes of cancer deaths. It is difficult to detect because it arises and shows symptoms in final stage. However, mortality rate and probability can be reduced by early detection and treatment of the disease. Best imaging technique CT imaging are reliable for lung cancer diagnosis because it can disclose every suspected and unsuspected lung cancer nodules [2]., However, variance of intensity in CT scan images and anatomical structure misjudgment by doctors and radiologists might cause difficulty in marking the cancerous cell [3]. Recently, to assist radiologists and doctors detect the cancer accurately computer Aided Diagnosis has become supplement and promising tool [4]. There has been many systems developed and research going on detection of lung cancer. However, some systems do not have satisfactory accuracy of detection and some systems still has to be improved to achieve highest accuracy tending to 100%. Image processing techniques and machine learning techniques has been implemented to detect and classify the lung cancer.

1.2. Problem Statement

Machine and deep learning methods are used widely in the medical and healthcare field for monitoring, detecting, classifying, and predicting diseases. Our system deals with CNN-SVM architecture which removes and eliminates useless information that negatively impacts accuracy. This is accomplished in the CNN architecture's pooling step, which is used to classify lung cancer in the fully connected layer using a modified SVM architecture.

1.3. Research Objective

In this project we are mainly addressing 3 objectives that are :

1.3.1. The primary objective of using machine learning (ML) for lung cancer detection is to enhance the accuracy and efficiency of identifying potential cases of lung cancer from medical data.

1.3.2. The objective of this project is to develop a hybrid Convolutional Neural Network (CNN) algorithm and Support Vector Machine (SVM) model.

1.3.3. CNN is known to have the best performance when using large amounts of data lacking in medical imaging due to several factors such as ethics and lack of well-labelled data.

1.4. Scope of the Project

The analysis and study of lung cancer/diseases has been the most intriguing investigation zone of medical experts from early days to the present day. To address this concern, a diagnosis system like this can only help diminish the odds of getting risk to human lives. The scope for this proposed system mainly works in medical field and it will be very useful to the medical experts.

CHAPTER 2
BACKGROUND WORK

2. BACKGROUND WORK

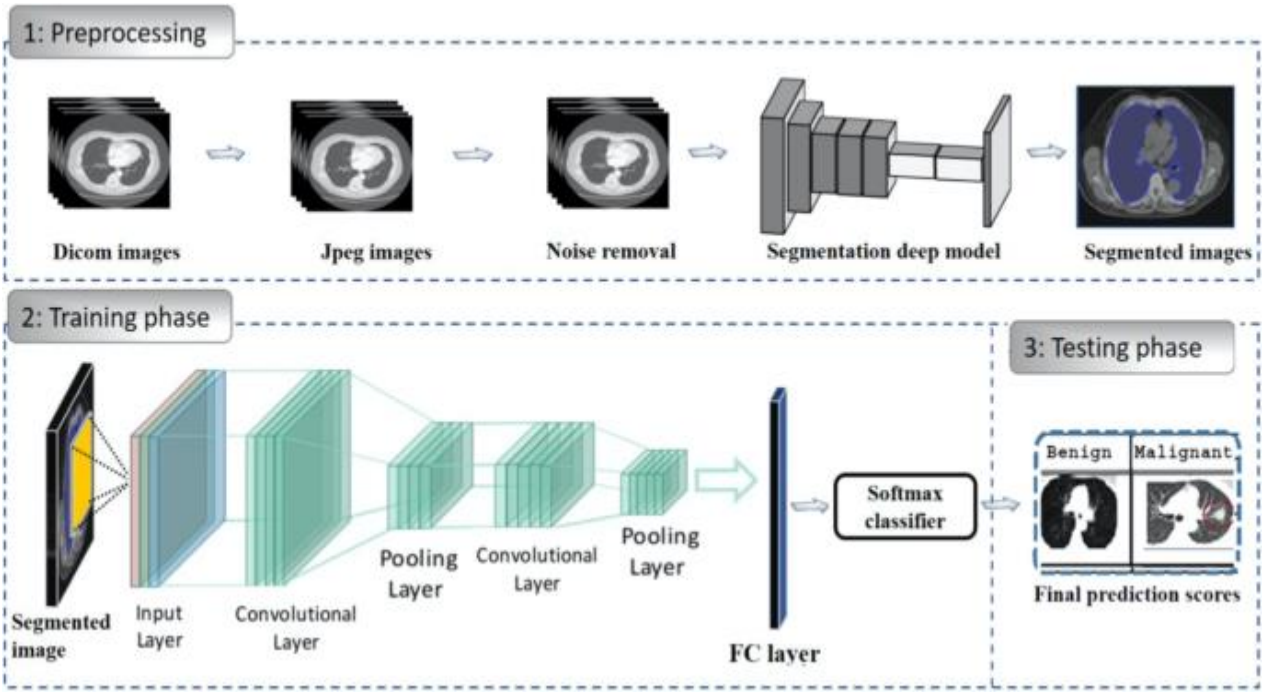
In this section we have studied various implementations of lung cancer detections & we summarized our findings that we concluded by researching & referencing various papers. They are as below:

Mokhled S. Al-Tarawneh (August,2012) [5]. Lung cancer is a disease of abnormal cells multiplying and growing into a tumor. Cancer cells can be carried away from the lungs in blood, or lymph fluid that surrounds lung tissue. Lymph flows through lymphatic vessels, which drain into lymph nodes located in the lungs and in the centre of the chest. Lung cancer often spreads toward the centre of the chest because the natural flow of lymph out of the lungs is toward the centre of the chest. Metastasis occurs when a cancer cell leaves the site where it began and moves into a lymph node or to another part of the body through the blood stream [6]. Several researchers has proposed and implemented detection of lung cancer using different approaches of image processing and machine learning. Aggarwal, Furquan and Kalra [7] proposed a model that provides classification between nodules and normal lung anatomy structure. The method extracts geometrical, statistical and gray level characteristics. LDA is used as classifier and optimal thresholding for segmentation. The system has 84% accuracy, 97.14% sensitivity and 53.33% specificity. Although the system detects the cancer nodule, its accuracy is still unacceptable. No any machine learning techniques has been used to classify and simple segmentation techniques is used. Therefore, combination of any of its steps in our new model does not provide probability of improvement. Jin, Zhang and Jin [8] used convolution neural network as classifier in his CAD system to detect the lung cancer. The system has 84.6% of accuracy, 82.5% of sensitivity and 86.7% of specificity. The advantage of this model is that it uses circular filter in Region of interest (ROI) extraction phase which reduces the cost of training and recognition steps. Although, implementation cost is reduced, it has still unsatisfactory accuracy. Sangamithraa and Govindaraju [9] uses K mean unsupervised learning algorithm for clustering or segmentation. It groups the pixel dataset according to certain characteristics. For classification this model implements back propagation network. Features like entropy, correlation, homogeneity, PSNR, SSIM are extracted using gray-level co-occurrence matrix (GLCM) method. The system has accuracy of about 90.7%. Image preprocessing median filter is used for noise removal which can be useful for our new model to remove the noise and improve the accuracy.

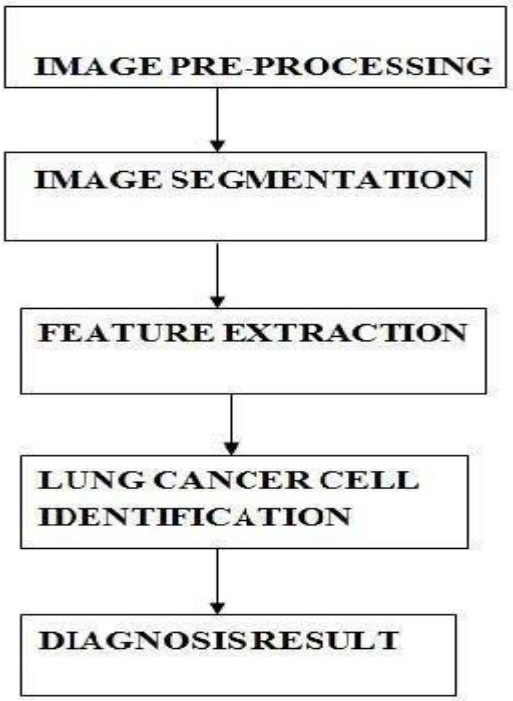
Roy, Sirohi, and Patle [10] developed a system to detect lung cancer nodule using fuzzy interference system and active contour model. This system uses gray transformation for image contrast enhancement. Image binarization is performed before segmentation and resulted image is segmented using active contour model. Cancer classification is performed using fuzzy inference method. Features like area, mean, entropy, correlation, major axis length, minor axis length are extracted to train the classifier. Overall, accuracy of the system is 94.12%. Counting its limitation it does not classify the cancer as benign or malignant which is future scope of this proposed model. Ignatious and Joseph [11] developed a system using watershed segmentation. In preprocessing it uses Gabor filter to enhance the image quality. It compares the accuracy with neural fuzzy model and region growing method. Accuracy of the proposed is 90.1% which is comparatively higher than the model with segmentation using neural fuzzy model and region growing method. The advantage of this model is that it uses marker controlled watershed segmentation which solves over segmentation problem. As a limitation it does not classify the cancer as benign or malignant and accuracy is high but still not satisfactory. Some changes and contribution in this model has probability of increasing the accuracy to satisfactory level.

When it comes to the detection of lung cancer using machine learning (Fig. 1)[12], a considerable amount of work has been done, a summary is provided (Table). Typically, a series of pre-processing steps using statistical methods and pretrained CNNs for feature extraction are carried out from several input sources (mostly images) to delineate the cancer region. Then, the extracted features are fed as input to several machine learning algorithms for classification of various lung cancer tasks such as the detection of malignant lung nodules from benign ones [13], the separation of a set of normalized biological data points into cancerous and non cancerous groups [14], and a basic comparative analysis of powerful machine learning algorithms for lung cancer detection [15].

The usage of machine learning techniques helps in early diagnosis and evaluation of lung nodules by processing CT images constructed through artificial intelligence methods. Such systems are called decision support systems that investigate the images through preprocessing, segmentation, feature extraction and classification process presented in (fig. 1).



Generalized machine learning framework for lung cancer prediction(fig.1) [12]



Flow chart (fig. 2)

Table 2.1: Comparison matrix table for various research papers studied

Reference	Author	Title	Year of Publishing	Results
[16]	Saba	Automated lung nodule detection and classification based on multiple classifiers voting.	2019	Accuracy- 96.410% Sensitivity-100%
[17]	Firmino et al.	Computer-aided detection (CADe) and diagnosis (CADx) system for lung cancer with likelihood of malignancy	2016	97.0 (Accuracy) 94.4 (Sensitivity) 7.04 (false positive)
[18]	S.M. Naqi, M. Sharif, I.U. Lali	A 3D nodule candidate detection method supported by hybrid features to reduce false positives in lung nodule detection.	2019	96.9 (Accuracy) 95.6 (Sensitivity) 97.0 (Specificity)
[19]	Asuntha and Srinivasan	Deep learning for lung Cancer detection and classification.	2019	95.62 (Accuracy) 97.93 (Sensitivity) 96.32 (specificity)
[20]	S.A. Khan, M. Nazir, M.A. Khan, T. Saba, K. Javed, A. Rehman	Lung nodule detection framework from computed tomography images using support vector machine	2017	96% (Accuracy) 97% (sensitivity) 96% (specificity) 97% (precision)
[21]	D. Kumar, A. Wong, D.A. Clausi	Lung nodule classification using deep features in CT images.	2015	5.01 (Accuracy) 83.35 (Sensitivity) 0.39 (false positive)

CHAPTER 3
RESULTS AND DISCUSSION

RESULTS AND DISCUSSION

The experiments are conducted on the proposed computer-aided diagnosis systems with the help of lung images obtained from the reputed hospital. This experimentation data consists of lung images. Those lung images are passed to the proposed CAD system. The diagnosis rules are then generated from those images and these rules are passed to the Support Vector Machine (SVM) for the learning process. After learning, a lung image is passed to the proposed CAD system. Then the proposed system will process through its processing steps and finally it will detect whether the supplied lung image is with cancer or not.

The last part is evaluating whether the proposed method has succeeded in enhancing the accuracy of prediction. Accuracy refers to the closeness of a measured value to a standard or known value. The accuracy of the result is very important in determining the best algorithm to use in the future. The higher the accuracy, the more excellent the results of the research are. The accuracy can be calculated based on:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

where TP refers to True Positives, TN refers to True Negative, FP refers to the False Positive, and FN refers to False Negative.

Similarly to find sensitivity, specificity and precision the following formulas are used.

$$\text{Sensitivity} = TP / (P+FN) \cdot 100\%$$

$$\text{Specificity} = TN /(TN+FP) \cdot 100\%$$

$$\text{Precision} = TP /(TP+FP) \cdot 100\%$$

Author	Method	Accuracy	Sensitivity	Specificity
Muzammil et al. [22]	Ensemble Learning Based Fusion	96.89%	95.59%	97.70%
Bansal et al. [23]	Deep3DSCcan	92.7%	87.10%	89.66%
Shah et al. [24]	Transfer Learning VGG16, VGG19	95.0%	84.0%	97.0%
Guo et al. [25]	Feature-Based and Optimized CNN	95.96%	97.10%	-

Jena et al. [26]	KNG-CNN	87.3%	97.85%	-

Table 3.1: Comparison of results using different methods

CHAPTER 4
CONCLUSION

CONCLUSION

In conclusion, This project introduces a hybrid CNN-SVM method to classify lung CT images into adenocarcinoma, large cell carcinoma, normal or squamous cell carcinoma. The aim was to achieve a higher level of accuracy, which is the goal of any computer-aided detection system. The method was applied to the Chest CT-Scan images dataset, a standard and publicly available cluster of CT images. The level of accuracy can be further improved by increasing the number of images utilized for the procedure. Additionally, X-ray, X-beam, and PET images can be interpreted by utilizing this method.

REFERENCES

[1] https://www.nccn.org/professionals/physician_gls/default.aspx. Accessed Jan. 13, 2020.

[2] 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.

[3] 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.

[4] 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.

[5] Mokhled S. Al-Tarawneh(August,2012),Lung Cancer Detection Using Image Processing Techniques.

[6] Non-Small Cell Lung Cancer, Available at: http://www.katemacintyrefoundation.org/ pdf/non-small-cell.pdf, Adapted from National Cancer Institute (NCI) and Patients Living with Cancer (PLWC), 2007, (accessed July 2011).

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] Ignatious, S., & Joseph, R. (2015) "Computer aided lung cancer detection system." 2015 Global Conference On Communication Technologies (GCCT), DOI: 10.1109/GCCT.2015.7342723.

[12] Saba T. Recent advancement in cancer detection using machine learning: systematic survey of decades, comparisons and challenges. J Infect Pub Health. 2020;13(9):1274–89.

[13] Saba T. Recent advancement in cancer detection using machine learning: systematic survey of decades, comparisons and challenges. J Infect Pub Health. 2020;13(9):1274–89.

[14] Abdullah DM, Abdulazeez AM, Sallow AB. Lung cancer prediction and classification based on correlation selection method using machine learning techniques. Qubahan Acad J. 2021;1(2):141–9.

[15] Abdullah DM, Abdulazeez AM, Sallow AB. Lung cancer prediction and classification based on correlation selection method using machine learning techniques. Qubahan Acad J. 2021;1(2):141–9.

[16] T. Saba Automated lung nodule detection and classification based on multiple classifiers voting Microsc Res Tech, 2019,pp. 1-9 10/1002/jemt.23326

[17] M. Firmino, G. Angelo, H. Morais, M.R. Dantas, R. Valentim Computer-aided detection (CADe) and diagnosis (CADx) system for lung cancer with likelihood of malignancy Biomed Eng Online, 15 (2016), p. 2.

[18] S.M. Naqi, M. Sharif, A. Jaffar Lung nodule detection and classification based on geometric fit in parametric form and deep learning Neural Comput Appl (2018), pp. 1-19

[19] A. Asuntha, A. Srinivasan Deep learning for lung Cancer detection and classification Multimed Tools Appl (2020), pp. 1-32

[20] S.A. Khan, M. Nazir, M.A. Khan, T. Saba, K. Javed, A. Rehman, *et al.* Lungs nodule detection framework from computed tomography images using support vector machine Microsc Res Tech (2019), 10.1002/jemt.23275.

[21] D. Kumar, A. Wong, D.A. Clausi Lung nodule classification using deep features in CT images 2015 12th Conference on Computer and Robot Vision, IEEE (2015), pp. 133-138.

[22] M. Muzammil, I. Ali, I. U. Haq, A. A. Khaliq and S. Abdullah, "Pulmonary nodule classification using feature and ensemble learning-based fusion techniques," IEEE Access, vol. 9, pp. 113415–113427, 2021.

[23] G. Bansal, V. Chamola, P. Narang, S. Kumar and S. Raman, "Deep3Dscan: Deep residual network and morphological descriptor based framework for lung cancer classification and 3D segmentation,"IET Image Process, vol. 14, no. 7, pp. 1316–1326, 2020.

[24] G. Shah, R. Thammasudjarit, A. Thakkinstian and T. Suwatanapongched, "Nodulenet: A lung nodule classification using deep learning," Ramathibodi Medical Journal, vol. 43, no. 4, pp. 11–19, 2020.

[25] Z. Guo, L. Xu, Y. Si and N. Razmjooy, “Novel computer-aided lung cancer detection based on convo lutional neural network-based and feature-based classifiers using metaheuristics,” International Journal of Imaging System and Technology, vol. 31, no. 4, pp. 1954–1969, 2021.

[26] S. R. Jena and S. T. George, “Morphological feature extraction and KNG-CNN classification of CT images for early lung cancer detection,” International Journal of Imaging System and Technology, vol. 30, no. 4, pp. 1324–1337, 2020.