International Journal of Current Research and Review
DOI: http://dx.doi.org/10.31782/IJCRR.2021.131711

Research Article

IJCRR
Section: Healthcare
ISI Impact Factor (2019-20): 1.628
IC Value (2019): 90.81
SJIF (2020) = 7.893

Development of Novel Technique to Detect and Validate Pulmo Malignancy during Early Stages

Dhanalakshmi R[1], Shree Harini R[2], Pravallika M[3], Sankar S[4]

[1,2,3,4]KCG College of Technology, Chennai, India.

ABSTRACT

Introduction: Lung carcinoma – Pulmonary disorders causes cancer-related death all over the world and in which majority due to cigarette smoking. With an increase in awareness about smoking being the major cause, other significant factors that play a vital role in causing the disease is unclear. There is no proper information among the public regarding the other symptoms which leads to identifying lung cancer in a later stage where it becomes incurable.

Aims: The proposed system helps in early diagnosis, effective treatment and helps in creating awareness about the danger of lung cancer in occasional or non-smokers too. This system is to predict Lung Cancer at an early stage and validate the results using a CT scan.

Methodology: An application that obtains user symptoms as input and prompts the user to upload a CT scan report of the lungs will be an efficient solution for early detection. This will aid in the early prognosis of the disease and effective treatment can be given. This application uses MATLAB to achieve its goal.

Results: Among the various methods analyzed, Naive Bayes achieved an accuracy of 95.24% which proves to be a better solution for detecting Lung Cancer

Conclusion: Thus, the proposed system has all the necessary features to detect lung cancer at an early stage thereby reducing the mortality rate and creating awareness among the public, of other parameters that are responsible for causing cancer.

Key Words: Pulmo Malignancy, Lung Cancer, Support Vector Machines, Prediction, Classification, Naive Bayes

INTRODUCTION

Modern medicine generates a great deal of information stored in medical databases. In today's world, every individual is facing growing health issues that need to be cured quickly. The useful information which is generated by using current medicine is stored in a medical database. With continually increasing lung cancer in patients due to the high intake of tobacco and puff, predicting cancer in patients at an early stage is a huge issue for clinicians to make decisions. Since it is considered a taboo in some countries people fear coming forward to diagnose the disease, the best place to find the occurrence of the disease is by applying machine learning concept to create the predictive model by using the data collected in the hospital regarding the patients affected by lung cancer to predict lung cancer.

In the 21st century, the most important cause of death and a hurdle in the longevity of the human race are NCDs. Non-communicable diseases(NCDs) are responsible for cancer and death worldwide. Lung cancer is an extensive reason for deaths globally. Lung cancer: 2,093,876 cases and death caused by lung cancer: 1,761,007 cases. Epidemiological progression in India is been huge in the past decades. In India, there is a sharp increase in chronic diseases and cancer cases have a steady impact on the illness. The outlook of the ancient and religious Indian medical system has only a few known facts about cancer which is changing fast and varied too.[6] In India, 70,000 new lung cancer cases are reported each year. A web-based application is developed to efficiently predict lung cancer using machine learning, acquire the factors that directly contribute to the disease and validate the results using a CT scan. This helps in early prognosis and effective treatment.

Corresponding Author:
Dhanalakshmi R, KCG College of Technology, Chennai, India.
Email: dhanalakshmisai@gmail.com

ISSN: 2231-2196 (Print) **ISSN:** 0975-5241 (Online)

Received: 19.01.2021 **Revised:** 10.02.2021 **Accepted:** 13.04.2021 **Published:** 12.09.2021

RELATED WORK

Predictive models are developed for cancer research which is effective and helps in making decisions precisely using techniques such as Bayesian networks, decision trees, artificial neural networks and support vector machines. Machine learning methods are proved to be effective in analyzing the progressive nature of cancer cells but a clear level of affirmation is needed to get implemented in regular clinical trials.

Kourou K et al. [10] presented an analysis report of the machine learning models in the field of cancer advancement. The different data samples and input features are used for different supervised machine learning techniques for the predictive models are reviewed. Krishnaiah et al. [9] suggested that naïve Bayes is the best model in predicting lung cancer in patients. If-then rule, decision trees and neural networks are later in the effectiveness of models. The results produced by decision trees are easy to read and understand. Decision trees are the only way to have a detailed analysis of patient profiles through the drill feature. But naïve bayes is the best as it can find all the important medical predictors compared to decision trees. To understand the relationship between attributes is very complex in neural networks. In order to enhance further, the prediction models can also be incorporated with other techniques such as association rules, clustering etc. Instead of categorical data, continuous data may be used. A large amount of unstructured data available in the health care industry can be mined. But the task is to define how to integrate text mining and data mining.

J Alam et al.[6] proposed a contrasting technique to recognize and predict lung cancer which gives better outcomes. SVM classifies a set of textural features derived from separated ROIs. The input image is used to find the tumour cells and their likely growth by this algorithm. Results are encouraging wherein cancer identification stands at 97% and prediction stands at 87% with the help of the results, doctors can identify whether the lung is carcinogenic or not. by using a genetic algorithm and deep neural network, the accuracy of the system can be improvised by having a huge image set and arrangement in LIGHT.

Hafan Yang et al.[7] explained that to have a lung cancer pathology report, a tissue sample is from the lung has to be taken through surgical biopsy. Replacing the pathology report with the clinical information of the patient will not put the health of a patient at risk. A correlation between pathology reports and the clinical information is derived using data mining techniques to give complete details on lung cancer pathologic staging diagnosis.

Kadir et al.[8] proposed a model to achieve performance in classification, CNN skilled with deep learning. The performance of AUC produced well and gave excellent accuracy with given data, but with independent data, it produced poor results. The following steps are involved to classify are segmentation, feature extraction, risk score regression and threshold. This algorithm dominates pattern recognition, segmentation and classification in considering medical and non-medical fields. Convolution Neural Network out-based Support Vector Machine method and previous state-of-art texture (radiomics) analysis of KAGGLE – which permits users to obtain and publish data sets.

Data science competition winners who used CNN trained data set using deep learning was done. Unlike AUC, the log loss function was used.[11] Therefore, the likelihood of cancer using the ct images was predicted. However, no nodules were found in acceptance and data for testing, so the automated reliable nodule finding step is challenging and complex for classification. Good results are generated in prediction based on size but the concern is size bias. In CADx (Computer-Aided Detection and Diagnosis system), nodule size will be enclosed as a part of nodule implicitly or explicitly. Therefore the efficiency of the CADx system is based on unmatched data and size-matched. Since a small data set is used, an SVM algorithm is applied. The AUC resulted in 0.70 when all benign images were included and when all malignant images and randomly selected benign images were included. In conclusion, if evaluation of system performance is after awareness whether the data contains smoker or non-smoker and current or prior history of malignancy is included CNN performs with high accuracy. From the present works that have been carried out, it is inferred that the recent trend involved in the prognosis of cancer is using machine learning. Naive Bayes is observed to give good accuracy. Using these algorithms efficient prediction of the people who are prone to be affected by lung cancer is done. The above researches give insight into the early prognosis of lung cancer. Using image processing, CT images of patients can be used to validate the presence of lung cancer in the individual. This will create awareness and helps to obtain the factors other than smoking that has a major effect of causing lung cancer in the population.

MATERIAL AND METHOD

The major cause of lung cancer is smoking but there is no rule that nonsmokers may never develop. Cancer cells can spread to any section of the body, metastasize to the lymph nodes. When the cells in the lung grow irregularly and are completely out of control and affects the nearby section and form a lump is referred to as lung cancer.[1]

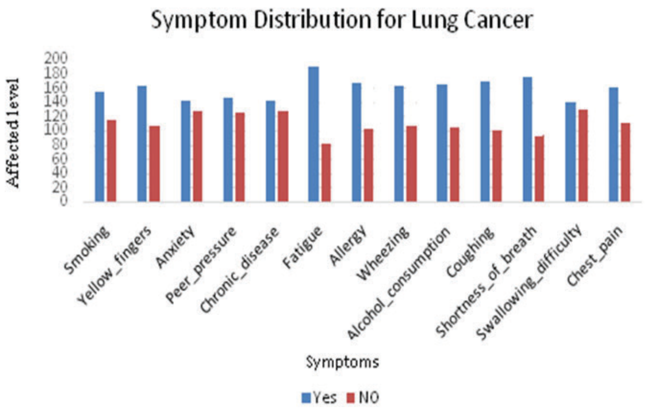


Figure 1: Naïve Bayes classification and symptom distribution.

Lung cancer may involve any section of the respiratory system and could start in any part of the lungs. Two types of cancer cells in lungs as follows small cell lung cancers (SCLC) which has the nature to develop very fast and non-small cell lung cancers (SCLC) which is less likely to spread in other section. The main cause of lung cancer is with the environment such as exposure to second-hand smoke, arsenic, asbestos, radioactive dust, or radon.[6] The chance of lung cancer increases with exposure to radiation at the workplace or anywhere. lung carcinogenic is greatly determined by the environment and genetic factors. the heritable contribution to the various histological subtypes is not known.[4] the indications are very general such as coughing, shortness of breath, wheezing, pain in chest and mucus in red colour when you cough. so people do not go for further examination to doctor in suspecting lung cancer. When cancer is detected, it would have invaded other sections already and few symptoms as in Fig 1. The preliminary identification of lung cancer is done by ct scan or x-ray. Furthermore, evaluation is needed to find the type of cancer cells and to the extent, it has spread[3]. The doctor can verify the reports and find the stage which is a mechanism to specify the size of cancer its spread.

Data extraction

In data mining algorithms, the accuracy of prediction is improved with the help of an accurate and specific dataset. Therefore in this investigation, understanding information on Lung malignancy infection is utilized. The data is collected from the website Online Lung Cancer prediction System that gets feedback from the user. In this database, 16 highlights of 310 individuals 207 of whom are not beneficial), which are considered as the fundamental benefactors of the illness, in the process of correlations and groupings are performed with lung. The precise outcome of the data mining process depends on the attributes which are considered in the investigation of disease. attribute considered are gender, age, yellow finger, anxiety, peer pressure, chronic disease, fatigue, allergy, alcohol consumption, smoking, pain in the chest, blood when coughing, shortness of breath, difficulty in swallowing, wheezing is taken to consider for identifying the lung cancer.

MATLAB[Matrix Laboratory] has the facility to perform data preprocessing, classification of data set using Naive Bayes.[5] The performance of this algorithm is analyzed using a confusion matrix. The presence or potentially the estimations of these parameters are firmly identified with the Lung malignancy data. feature reduction, class Currently available lung cancer CT image scans are obtained from an online resource: The Cancer Imaging Archive (TCIA). The images are preprocessed using feature selection and fine tuning[13]. Furthermore, a convoluted neural network algorithm is applied to the images and trained to classify benign and malignant tumours. The proposed workflow is given in Fig 2

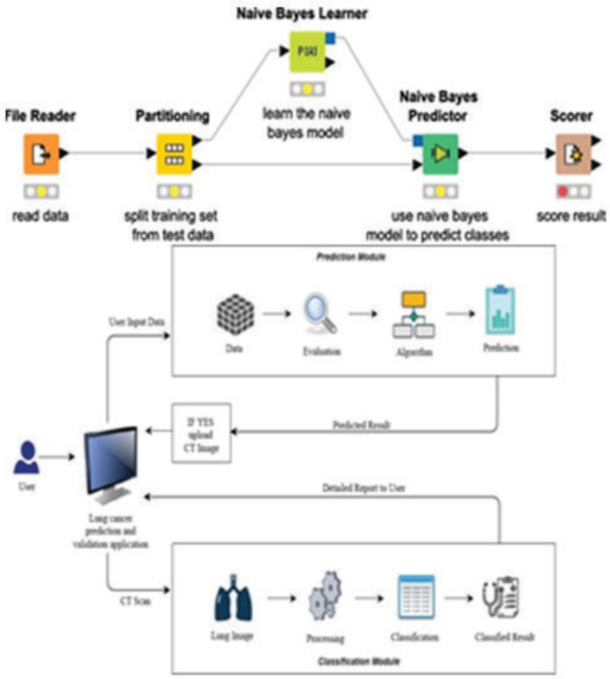


Figure 2: Proposed workflow diagram.

Currently, available lung cancer CT image scans are obtained from an online resource: The Cancer Imaging Archive (TCIA). The images are preprocessed using feature selection and fine tuning[13]. Furthermore, a convoluted neural network algorithm is applied to the images and trained to classify benign and malignant tumours. The proposed workflow is given below in Fig 2.

MODULE DESCRIPTION

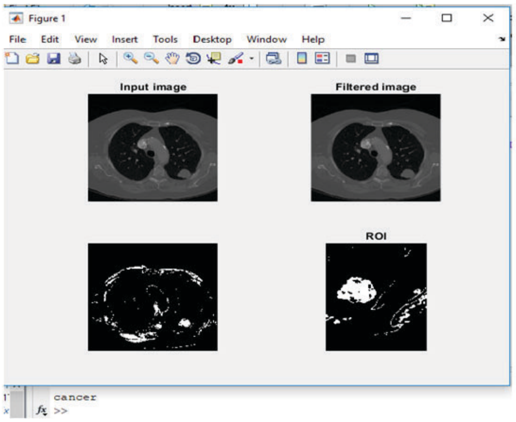
The proposed workflow in Fig 1 consists of two important modules as Lung Cancer Prediction and Image Classification.

Lung Cancer Prediction

In this phase, the obtained information from the user is first processed and the Naive Bayes algorithm is applied. Based on it, the trained system gives out the result which is either positive or negative. If it is positive, the patient is likely to be affected by lung cancer, else fortunate with the absence of tumour.

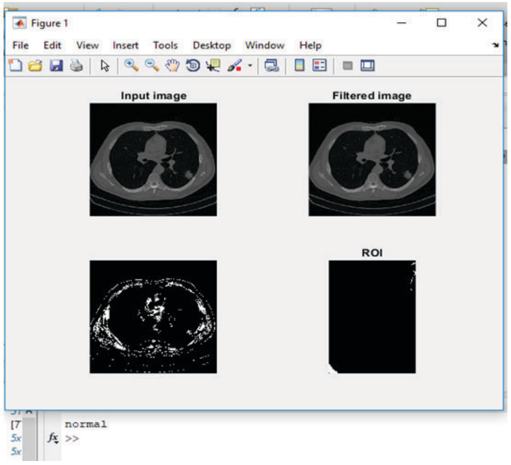
Image Classification

The primary steps involved in image classification is image preprocessing, feature extraction, selection of training samples, identifying the appropriate classification algorithm, the processing involved after classification and accuracy estimation. The client data in regards to the symptoms of lung cancer will be the initial step. The application accumulates the information and is passed to the prediction module. The prediction module comprises four steps such as information preparing, assessment, testing with model and foreseeing results. The outcome is then displayed to the client. The user is given an option to submit the CT image of the lungs. The image is then fed into the classifier which processes the image and applies a classification algorithm to segregate the tumour either as benign or cancerous. The classified outcome is produced as an answer to the client.



```
cancer
fx >>
```

Figure 3: Image classification.



```
normal
fx >>
```

Figure 4: Lung image with Malignancy.

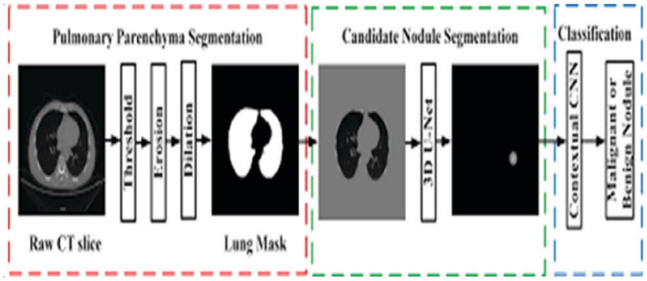

Pulmonary Parenchyma Segmentation
Threshold
Erosion
Dilation
Raw CT slice
Lung Mask
Candidate Nodule Segmentation
3D U-Net
Classification
Contextual CNN
Malignant or Benign Nodule

Figure 5: Lung image with No Malignancy.

EXPERIMENTAL ANALYSIS

The below table describes the performance analysis of the algorithm analyzes the accuracy score of the algorithms.

Table 1: Experimental Results Comparison

NAIVE BAYES	95.24%
KNN	89.90%
LOGISTIC REGRESSION	88.53%
SVM - LINEAR	88.04%
SVM - RADIAL	85.87%

The typical efficiency of the current system is 90.2% and for the proposed framework, the higher precision accomplished is 95.24% utilizing Naive Bayes and practically 88% for the other algorithms. The performance is analyzed and gives the outcome for higher accuracy in the prediction of Lung Cancer.[7]

CONCLUSION

In this paper, an ingenious multi-layered way to combine prediction and classification methods to develop a cancer risk prediction is suggested. malignant growth has turned into the main cause of death all over the world. The best method to diminish cancer deaths is to detect it earlier. Individuals maintain a strategic distance from malignant growth screening because of the cost associated with stepping through a few examinations for determination. This forecast framework may give a simple and practical route for screening disease and may assume an essential job in the prior finding process for various kinds of malignant growth and give a compelling preventive system. Furthermore, the validation technique confirms the predicted results. This system gives direction for the specialists to target specific treatment for patients depending on the detailed historical record of the patients available in the medical clinics.

ACKNOWLEDGEMENT

We thank our colleagues and Physicians who provided insight and expertise that greatly assisted the research, although they may not agree with all of the interpretations/ conclusions of this paper. We are also grateful to authors/ editors/publishers of all those articles, journals and books from where the literature for this article has been reviewed and discussed

Conflict of Interest and Funding

There is no conflict of interest

Contribution of Authors

1. R Dhanalakshmi for the idea and structuring this paper
2. M Thenmozhi for literature review
3. M Pravellika and Shree Harini for Implementation and Results

REFERENCES

1. Broom A, Kenny K, Bowden V, Muppavaram N, Chittem M. Cultural ontologies of cancer in India. Critical Public Health. 2018 Jan 1;28(1):48-58.
2. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: Cancer J. Clin. 2018 Nov;68(6):394-424.
3. Carr SR, Akerley W, Cannon-Albright LA. Genetic Contribution to Nonsquamous, Non–Small Cell Lung Cancer in Nonsmokers. J Thorac Oncol. 2018 Jul 1;13(7):938-45.
4. Gnagnarella P, Caini S, Maisonneuve P, Gandini S. Carcinogenicity of high consumption of meat and lung cancer risk among non-smokers: a comprehensive meta-analysis. Nutrit Canc. 2018 Jan 2;70(1):1-3.
5. Alam J, Alam S, Hossan A. Multi-stage lung cancer detection and prediction using multi-class svm classified. In2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2) 2018 Feb 8 (pp. 1-4). IEEE.
6. Yang H, Chen YP. Data mining in lung cancer pathologic staging diagnosis: Correlation between clinical and pathology information. Expert Syst Applications. 2015 Sep 1;42(15-16):6168-76.
7. Kadir T, Gleeson F. Lung cancer prediction using machine learning and advanced imaging techniques. Translational lung cancer research. 2018 Jun;7(3):304.
8. Krishnaiah V, Narsimha G, Chandra DN. Diagnosis of lung cancer prediction system using data mining classification techniques. Int J Sus The. 2013 Apr;4(1):39-45.
9. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. CSBJ. 2015 Jan 1;13:8-17.
10. Sumathipala Y, Shafiq M, Bongen E, Brinton C, Paik D. Machine learning to predict lung nodule biopsy method using CT image features: A pilot study. Computerized Medical Imaging and Graphics. 2019 Jan 1;71:1-8.
11. Wakelee HA, Chang ET, Gomez SL, Keegan TH, Feskanich D, Clarke CA, Holmberg L, Yong LC, Kolonel LN, Gould MK, West DW. Lung cancer incidence in never-smokers. *J* Clin Oncol. 2007 Feb 10;25(5):472.
12. Johnson, M., Dhanalakshmi, R. Predictive Analysis based Efficient Routing of Smart Garbage Bins for Effective Waste Management.
13. Booma, P. M., Prabhakaran, S., & Dhanalakshmi, R. (2014). An Improved Pearson's Correlation Proximity-Based Hierarchical Clustering for Mining Biological Association between Genes. The Scientific World Journal, 2014.