International Journal of Scientific Research and Engineering Development-– Volume 5 Issue 3, May-June 2022

Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Lung Cancer Detection using Artificial Intelligence and Machine Learning

Abhishek*, Charu Jain**, Ankit Garg***
*(Department of Computer Science and Technology, Amity University, Haryana
Email: abhitherage16@gmail.com)
**(Department of Computer Science and Technology, Amity University, Haryana
Email: cjain@ggn.amity.edu)
***(Department of Computer Science and Technology, Amity University, Haryana
Email: agarg1@ggn.amity.edu)

--************************----------------------------------

Abstract:

Lung cancer has been termed as the most fatal of all types of cancer and is affecting thousands of people around the world every year. If not detected in its initial phase, rate of survival of patients is grim. It is the major death causing form of cancer. Nearly 70,000 cases every year are found in India. This makes it a serious concern for finding ways to diagnose it in its initial state so as to provide required treatment to the patient thus saving his life. Lung cancer detection using machine learning approach has been an area of interest for researchers to find a solution to this challenging problem. CNN (Convolution Neural Networks) models proved as more advanced tool to diagnose this disease giving higher accuracy and performance. This paper presents a system to categorize tumors found in the lungs as malignant or benign using CNN model so that the patient can be treated accordingly. Accuracy achieved by this model is 99% making it more efficient than the existing systems. CT scan images of lung cancer being used as the dataset from online sources.

Keywords — *Lung Cancer Detection , Artificial Intelligence , Machine Learning , Deep Learning , SCLC and NSCLC.*

--************************----------------------------------

I. INTRODUCTION

Lung cancer is the most found types in both genders and contributes 25% to deaths due to all forms of cancer.[1] Primary causes leading to death by lung cancer occurs 80% by smoking. Non-smokers gets affected by coming in the influence of exposed radon, second-hand smoke, air pollution, diesel exhaust and other harmful chemicals are major causes.[2]Numerous techniques like X-ray, CT scan, biopsy, are implemented to detect cancer cells in lungs. During the process of biopsy, microscopic histopathology slides are evaluated by trained pathologists to carry out the diagnosis.[3],[4],[5], and based on the findings classify the cancer types. This process takes a considerable amount of time. If not detected in the early stages, the cancer cells spreads at an alarming rate resulting the decreased chances of survival of patients.

Lung cancer is divided into two major categories: Non-small cell lung cancer (NSCLC) and Small cell lung cancer (SCLC).

Out of all cancer cases, it is found that nearly 80% lung cancer are NSCLC but this is further divided into various types that are adenocarcinoma, squamous cell carcinoma, and large cell carcinoma. Adenocarcinomas usually begins in the cells continuously giving out discharge substances such as mucus. This cancer type occurs majorly to people who have been smoking or used to smoke but it is also equally found in non-smokers also. Female and young people are more affected by this cancer. Squamous Cell carcinoma begins from the squamous cells present in the aviation routes in lungs. Found majorly in people having a background of smoking. Large cell carcinoma can be found in any part of the lungs. Usually, it spreads swiftly which makes the diagnosis tough. Small cell lung cancer contributes to 15% of the total cancer patients, often referred as oat cell carcinoma. This type generally develops faster than NSCLC. Chemotherapy and radiation therapy works well with this type because the malignant grows rapidly. But there are chances that this malignancy will come back sooner or later.

Machine Learning (ML) is termed as the sub-part of Artificial Intelligence (AI) which facilitates the machines to learn irrespective of any explicit programming through set of data enabling them to learn a particular method through experience. According to the earlier research, most authors preferred use of x-rays and CT scan images coupled with machine learning algorithms namely Support Vector Machines (SVM), Random Forest(RF), Bayesian Networks (BN), and Convolutional Neural Networks (CNN) for the purpose of cancer detection. This paper projects a system developed to detect lung nodules in the CT scan images dataset taken from online sources and classify the results as malign or benign. Deep learning algorithms are implemented using Convolution Neural Networks to prepare the model and tested across various parameters namely accuracy, precision, recall value, f-measure, specificity. The model returned training accuracy and validation accuracy of 99%.

II. LITERATURE REVIEW

Carrillo et al.[6] presented a framework for combining five multi-scale and one multi-omic modalities, namely RNA-Seq, miRNA-Seq, whole-slide imaging, copy number variation, and DNA methylation, to conduct a study employing multi-scale and multi-omic cancer data. Late fusion strategy and machine learning techniques were used as technology. For each modality, the model was trained separately, and output was obtained by fusing gains in an ascending order. After incorporating all modalities, the final model had a precision of 96.82 percent. The findings clearly illustrate that including the multi-scale and multi-omic character of cancer data might increase the performance of single-modality decision support systems, thereby assisting in the diagnosis process.

Cancer, according to Md. Alamin et al.[7], is a deadly disease caused by a combination of genetically present disease and anomalies in the human body. The most important stage in determining the optimal treatment for the patient is histopathological detection. If found early, the death rate can be considerably reduced. A hybrid ensemble feature extraction model is employed in the proposed system to detect the presence of lung and colon cancer. Deep feature extraction and ensemble learning techniques used to histopathology (LC2500) lung datasets made this achievable. The model was found to have a 99 percent accuracy in detecting lung cancer, making it suitable for clinical trials to improve the detection of this devastating disease.

KnowSeq, according to Castillo et al.[8], is intended to be a powerful and scalable modular software that focuses on the automation and assembly of bioinformic tools with modern functionalities. It was a unified environment designed to undertake extensive gene analysis in order to diagnose a specific disease. Raw files from well-known platforms can be used in the process, or it can be passed by the users themselves. The most representative genes in the stated problem are chosen using a set of advanced algorithms. KnowSeq generates a full report on the entire task automatically. To determine the efficacy of this strategy, researchers looked at biclass breast cancer and multiclass lung cancer. The approach had a 95 percent accuracy rate.

According to A. Rehman et al.[9], lung cancer is one of the leading causes of death worldwide, with nearly five million cases recorded each year; nevertheless, early detection, if possible, can improve the diagnosis process. Images from a computed tomography (CT) scan showed to be the most useful for diagnosing lung infections. A cancer diagnosis technique was developed employing machine learning techniques such as feature extraction, fusion using LBP (Local Binary Pattern) and DCT (Discrete Cosine Transform), SVM (Support Vector Machine), and K-nearest neighbour in the suggested system. In comparison to other state-of-the-art techniques, the model achieved an accuracy of 93 percent for SVM and 91 percent for K-nearest neighbours.

Bhandary et al.[10] propose a modified AlexNet (MAN) to evaluate lung anomalies in evaluated pictures. The two types of images considered are chest X-rays and lung CT scans. On these two photo datasets, the proposed MAN is tested individually. During the initial diagnostic phase, the thoracic X-ray is evaluated as normal, and the pneumonia class is decided. When compared to the other strategies presented in this study, the proposed deep learning method has a 96 percent accuracy rate.

In this study, Muhammad Imran Faisal et al [11] aim to examine the discriminative intensity of a few predictors in order to increase the productivity of lung cancer detection through their manifestations. Support Vector Machine (SVM), C4.5 Decision Tree, and Multi-Layer Classifiers are among the several classifiers.On a benchmark dataset obtained from the repository at UC Irvine, Perceptron, Neural Network, and Naive Bayes (NB) were evaluated.The exhibition was also contrasted and prominent ensembles, such as Random, were featured.Majority Voting and the Forest. Gradient-boosted Tree outperforms other trees in terms of execution.All other persons worked as group classifiers and achieved a 90% accuracy rate.

For lung cancer classification, the authors Jay Kumar Raghavan Nair, MD et al [12] employed logistic regression as a machine learning classifier. They employed a lung cancer imaging characteristics data collection. A total of fifty patients are included in the data set. They discovered a respectable accuracy score of 71 to 78 percent. All other persons worked as group classifiers and achieved a 90% accuracy rate.

After conducting literature review the following research gaps are identified :

- The determination strategies are tedious, as in sickness expectation, highlight choice techniques assume a key part in cancer forecast.
- The customary techniques are complicated and precision is additionally low.
- The extraction process is not easy to perform on the dataset and the malignant growth cell extraction process is tedious.
- The customary technique neglects to distinguish cancers in the beginning phases and with least sizes.

III. RESEARCH OBJECTIVES

The primary goal of this study is to develop a computer-aided detection system for lung cancer nodules in chest CT images. This study aims to create a CNN system with the following features:

- High accuracy
- Low false positive and true positive detection rates
- Faster classification time
- Increased sensitivity and specificity

IV. METHODOLOGY

A. System Architecture

The steps performed in system architecture are described below :

1. At first , CT scan image is taken from the dataset available online to test the presence of cancer.
2. Image processing techniques like feature extraction are performed on the input image so that it becomes suitable for model to analyze it. Processing of image consists of the various steps performed to format the images before using them for training the

model. Some commonly used preprocessing techniques are resizing, orienting, color corrections, denoising, segmention, morphology etc.

3. After pre-processing the image is sent to the CNN model architecture which runs on machine learning algorithms. CNN is a feed-forward type of network in which the link between neurons is demonstrated by the architecture of human cortex. It consists of layers namely :
   - Convolutional layer,
   - ReLU layer,
   - Pooling layer,
   - Fully connected layer
4. The input image is predicted as cancerous or non-cancerous by the model. Prediction is the output of trained algorithm using a historical dataset. Naïve Bayes is a simple yet powerful algorithm that is used for prediction purpose. The model contains two types of probabilities that are calculated using trained data.
5. Performance statistics are displayed along with confusion matrix graph which helps to study the results. It displays the various parameters like accuracy, precision, F1 score, recall value, specificity that aids to analyzing the performance of the model.

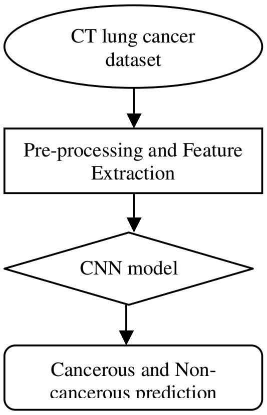

CT lung cancer dataset
Pre-processing and Feature Extraction
CNN model
Cancerous and Non-cancerous prediction

Figure 1. Model Architecture

B. Implementation of model

The implementation of the project is elaborated as :

- First the dataset for training and testing of the model is implemented which consists of 550 CT scan images in a single folder for cancer prediction.
- After that the necessary libraries for creating the model are imported.
- The model is created using Python language so libraries under python like sklearn, pandas, PIL, numpy, matplotlib, tensorflow, keras are imported.
- Now the labels are retrieved from the images and they are resized to (224,224) to make them similar using image resizing method.
- The square measure of the images is stored in an numpy array for later analysis.
- Dataset is classified into two sections : Testing and training.
- For constructing the model, Keras library is used to form various layers of CNN.
- FPCM (Fuzzy Possiblistic C-means) is used to extract the lung region from the input image from which the segementation of lung region takes place which is then analysed and tested for evaluation.

C. Dataset

CT scan images are used as test and train data taken from online platform. The images are divided into two categories Cancerous and Non-cancerous for test and train data respectively. Total CT scans images of both type counts to 550. This dataset comprises of the actual CT scan images which makes it suitable for testing with the developed model to get real time results and increase the efficiency of model to make it useful in medical analysis in future. The images are high resolution quality which aids in the smooth extraction of regions and processing the required parameters to give final results.

D. Input Image

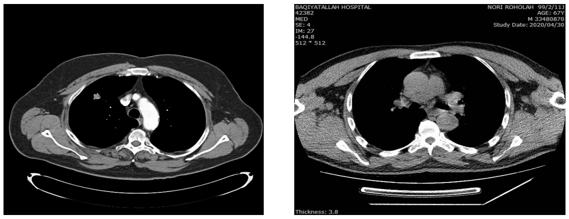


Figure 2. Input Image – Cancerous and Non-cancerous CT scan

Figure 2 represents the sample of cancerous and non-cancerous CT scan images being used as test data to train the model. Around 250 Cancerous images and 250 non-Cancerous images are available in the dataset for testing and can be passed as an input.

E. Technologies Used

To carry out the research work Python language has been used with Jupyter Notebook.

Python is a high-level programming language which is used to develop models by deploying machine learning algorithms. It is the most suitable form to develop models using artificial intelligence and machine learning techniques. It has a large collection of libraries like pandas, numpy, keras, tenserflow, matplotlib, opencv that contains predefined functions for various purposes. Developed by Guido van Rossum, it can be applied in the fields of web development, software development, mathematics, and system scripting.[15]

Convolutional neural network is a type of Deep Learning algorithm developed to work on images and videos. This algorithm accepts input in form of images and then extracts features from it and learns on its own to classify them according to these features. The algorithm draws its usefulness from cortex which signifies how a human brain works. CNN model comprises of two sections : feature extraction and classification. Feature extraction is performed by application of numerous filters and layers to extract the information and features and then send to the second section which is classification where the extracted features are classified on the basis of the target variable.[16]

CNN model as shown in figure 4 comprises of :

- Input layer
- Convolution layer and activation function
- Pooling layer
- Fully connected layer

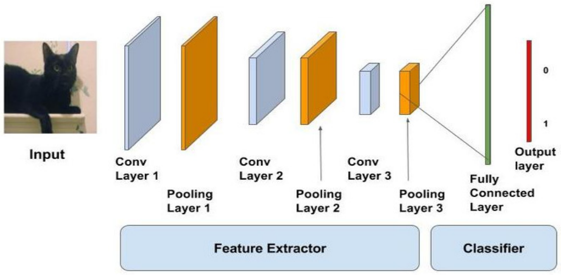


Figure 3. CNN model representation

F. Parameters Used

The model has been tested across the various parameters like Specificity, Accuracy, Precision, F1 score, Recall value.

Specificity is termed as the measure of true positive predicted accurately by the model.

Specificity = TN/ TN + FP (1)

Precision acts as an indicator of performance for a machine learning model. Precision is calculated by dividing true positives to the sum of true positives and false positives.

Precision = TP/TP + FP (2)

Recall is the ratio of number of true positives samples that are classified accurately to the total number of positive samples.

Recall = TP/TP + FN (3)

F1 score clearly depicts the percentage obtained by the model after making the correct predictions.

F1 = TP/TP + ½(FP + FN) (4)

Accuracy is used as a parameter to determine the most suitable model after carrying out the relationships and patterns among variables on the basis of input data.

Accuracy = TP + TN/ TN+TP+FN+FP (5)

V. ANALYSIS OF RESULTS

Edge detection is performed on the input images and the results are compared with the non-cancerous data to clearly distinguish the nodule present in the lungs. This is the most effective method to detect the presence of tumors in lungs as this method is able to detect even very small size anomalies in the region. The input image is split into two halves namely vertical edges and horizontal edges. After that to detect the edges the images are converted to grayscale image. The grayscale image is convolved in the form of matrix of 6 X 6 with a 3 X 3 filter. After convolving the resulting matrix obtained is 4 X 4. Now, to calculate the second component of the resulting 4 X 4 matrix the channel is moved further one stage to the right and again the amount of component is calculated. Features applied on the test image are :

- Padding
- Cushioning
- Convolutions over volume

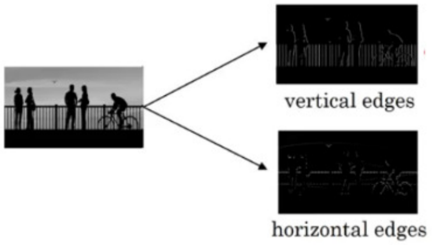


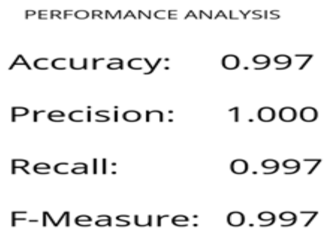
Figure 4. Splitting of image for Edge Detection

The formula for yield can be summed up as :

- Input : n X n X nc
- Channel : f X f X nc
- Cushioning : p
- Step : s

Yield = [(n+2p-f)/s+1]X[(n+2p-f)/s+1]X nc (6)

After testing the models for numerous test data images the results obtained were quite satisfactory. The proposed model is capable to identify the CT scan images as cancerous or non-cancerous with a training accuracy of 99% and validation accuracy of 99%.The specificity obtained is 99% proves that less number of false positives were detected as shown in figure 5.



PERFORMANCE ANALYSIS

Accuracy:	0.997
Precision:	1.000
Recall:	0.997
F-Measure:	0.997

Figure 5. Performance Analysis

Figure 6 represents the confusion matrix values obtained for cancerous and non-cancerous data. It clearly shows that the model developed is quite efficient to detect cancer nodules present in the lungs.

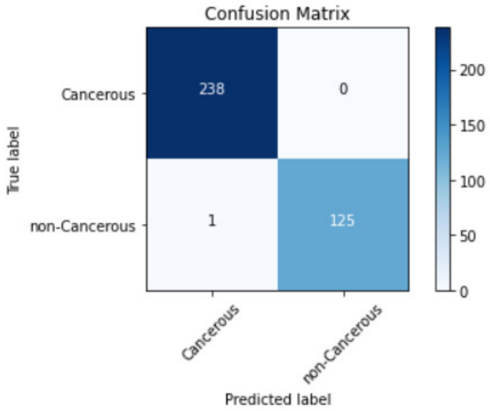


Figure 6. Confusion Matrix

CONCLUSIONS

Early detection of cancer can efficiently increase the survival rate of patients. The proposed proves its usefulness in the field of medical sciences and will aid doctor in early diagnosis of the disease. It is able to fill the research gas found in the previous systems making it superior over them. The model is able to classify the lung images as cancerous or non-cancerous with an accuracy of 99% and least false positive rate was obtained. CNN proved to be the most efficient method over the other techniques like SVM, K-nearest neighbours, Random forest. This model can be implemented to real world medical imaging machines to aid the doctors to examine the disease with improved and satisfactory accuracy and help the patients to get proper treatment in time. Further, this will also help to find the anamolies and challenges that will come across which will make a way to improve this model and equip it with the required functionalities. With the advancement of technology, this is an important research area where new tools and techniques will be available in future that will make the process of detection much efficient and provide more and more information of the detected lung nodules.

This model should be tested for large datasets to examine the accurately examine the shape and size of the lung nodule. The precision of the model can be further improved by using 3D Convolutional Neural Network. There is also a need to work on the foggy clinical pictures having spots to distinguish the growth with proofs. Algorithm can also be updated by adding more parameters like extracting and segmenting the lung nodule along with shape, size, and color and various cases can be monitored to declare an approximate rate with which these cancer cells spread in the human body giving the doctors an estimate about seriousness of the disease and treat the patient accordingly before its too late.

REFERENCES

[1] (2020) "*American Cancer Society, Lung Cancer Statistics. [Online]*". Available: https://www.cancer.org/cancer/lung-cancer/about/key-statistics.html

[2] (2019) "*American Cancer Society, Lung Cancer Causes. [Online].*" Available: https://www.cancer.org/cancer/lung-cancer/causes-risks-prevention/what-causes.html

[3] G. A. Silvestri, et al. "*Noninvasive staging of non-small cell lung cancer: ACCP evidence-based clinical practice guidelines (2nd edition).*" Chest vol. 132, 3 Suppl (2007): 178S-201S. doi:10.1378/chest.07-1360.

[4] W. D. Travis, et al. "International *association for the study of lung cancer/American thoracic society/European respiratory society international multidisciplinary* classification of lung adenocarcinoma." Journal of thoracic oncology: official publication of the International Association for the Study of Lung Cancer vol. 6, 2 (2011): 244-85. doi:10.1097/JTO.0b013e318206a221

[5] L. G. Collins., C. Haines, R. Perkel & R. E. Enck. *"Lung cancer: diagnosis and management."* American family physician vol. 75, 1 (2007): 56-63.

[6] Carrillo-Perez, Francisco, Juan C. Morales, Daniel Castillo-Secilla, Olivier Gevaert, Ignacio Rojas, and Luis J. Herrera. 2022. "Machine-Learning-Based Late Fusion on Multi-Omics and Multi-Scale Data for Non-Small-Cell Lung Cancer Diagnosis" *Journal of Personalized Medicine* 12, no. 4: 601.

[7] Md. Alamin Talukder, Md. Manowarul Islam, Md Ashraf Uddin, Arnisha Akhter, Khondokar Fida Hasan, Mohammad Ali Moni,Machine learning-based lung and colon cancer detection using deep feature extraction and ensemble learning, Expert Systems with Applications, Volume 205,2022,117695,ISSN 0957-4174.

[8] Castillo-Secilla, D.; Gálvez, J.M.; Carrillo-Perez, F.; Verona-Almeida, M.; Redondo-Sánchez, D.; Ortuno, F.M.; Herrera, L.J.; Rojas, I. KnowSeq R-Bioc package: The automatic smart gene expression tool for retrieving relevant biological knowledge. *Comput. Biol. Med.* 2021, *133*, 104387.

[9] Rehman, M. Kashif, I. Abunadi and N. Ayesha, "Lung Cancer Detection and Classification from Chest CT Scans Using Machine Learning Techniques," *2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA)*, 2021, pp. 101-104, doi: 10.1109/CAIDA51941.2021.9425269.

[10] Bhandary, A., Prabhu, G. A., Rajinikanth, V., Thanaraj, K. P., Satapathy, S. C., Robbins, D. E., & Raja, N.S. M. (2020). Deep-learning framework to detect lung abnormality–A study with chest X-Ray and lung CT scan images. Pattern Recognition Letters, 129, 271-278.

[11] Muhammad Imran Faisal, Saba Bashir, Zain Sikandar Khan, Farhan Hassan Khan," An Evaluation of Machine Learning Classifiers and Ensembles for Early Stage Prediction of Lung Cancer" 2020.

[12] Jay Kumar Raghavan Nair, Umar Abid Saeed, Connor C. McDougall, Ali Sabri, Mmed, Bojan Kovacina, B. V. S. Raidu, Riaz Ahmed Khokhar, Stephan, Vera Hirsh, Chankowsky Jeffrey, Leon C. Van Kempen, and Jana Taylor, "Radiogenomic Models Using Machine Learning Techniques to Predict EGFR Mutations in Non-Small Cell Lung Cancer", https://doi.org/10.1177/0846537119899526 , The Author(s) 2020.

[13] S. Mehmood *et al.*, "Malignancy Detection in Lung and Colon Histopathology Images Using Transfer Learning With Class Selective Image Processing," in *IEEE Access*, vol. 10, pp. 25657-25668, 2022, doi: 10.1109/ACCESS.2022.3150924.

[14] Sakib, SM Nazmuz. "Research Proposal: Lung Cancer Prediction and Classification using Machine learning Models." *Authorea Preprints* (2022).

[15] https://www.w3schools.com/python

[16] https://www.analyticsvidhya.com/blog/2021/08/beginners-guide-to-convolutional-neural-network-with-implementation-in-python