IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, July 2022
Impact Factor: 6.252

Lung Cancer Detection using Machine Learning

Madhushree A[1], Harshitha Nayaka YS[2], Chandrika M[3], Madangowda HS[4], Mrs. Pallavi J[5]
Students, Department of ECE[1,2,3,4]
Assistant Professor, Department of ECE[5]
Vidya Vikas Institute of Engineering and Technology, Mysuru, Karnataka, India

Abstract: *Cancer is a disease in which cells in the body grow out of control. When cancer starts in the lungs, it is called lung cancer. Lung cancer begins in the lungs and may spread to lymph nodes or other organs in the body, such as the brain. Cancer from other organs also may spread to the lungs. When cancer cells spread from one organ to another, they are called metastases.*

Keywords: Lung cancer detection, python code, Kaggle, Keras library, Google collab

I. INTRODUCTION

Cancer is a lethal disease that is frequently brought on by the accumulation of hereditary disorders and many pathological alterations. Cancerous cells are abnormal growths that can appear anywhere on the body and pose a threat to life. To determine what might be helpful for its treatment, cancer, also known as a tumour, must be promptly and accurately discovered in the early stages. Even while each modality has its own unique considerations, some of the major causes of mortality include difficult histories, inadequate diagnoses, and inadequate treatment. When you inhale, air enters through your mouth or nose and travels through your trachea to your lungs (windpipe). The trachea splits into bronchi, which pass into the lungs and further split into smaller bronchi. These divide into bronchioles, which are smaller branches. The tiny air sacs known as alveoli are located at the end of the bronchioles. When you inhale air, the alveoli take oxygen into your blood and eliminate carbon dioxide when you exhale.

The major jobs of your lungs are to take in oxygen and expel carbon dioxide. The cells lining the bronchi and other areas of the lung, such as the bronchioles or alveoli, are where lung malignancies generally begin. The pleura, a slender layer of lining, encircles the lungs. Your pleura shields your lungs and aids in their movement back and forth against the chest wall as you breathe.

The diaphragm, a slender, dome-shaped muscle, divides the chest from the belly under the lungs. The diaphragm contracts and expands while you breathe, propelling air into and out of the lungs.

The purpose of is to study and analyse methods for lung cancer diagnosis using machine learning. The study demonstrates how machine learning utilising supervised, unsupervised, and deep learning approaches is used to aid in the detection and treatment of cancer. The performance of several state-of-the-art approaches is compared using benchmark datasets for accuracy, sensitivity, specificity, and false-positive rates.

Artificial neural networks do incredibly well in machine learning. Artificial neural networks are used to classify words, audio, and images among other things. Different forms of neural networks are employed for various tasks. For example, to predict the order of words, recurrent neural networks— more specifically, an LSTM—are used. Similarly, to classify images, convolution neural networks are employed. We're going to create the fundamental building element for CNN in the inputs are images are known as convolutional neural networks (CNN or ConvNets). They are used to conduct object detection within a frame, evaluate and categorise photos, and group images based on similarities. Convolutional neural networks (ConvNets or CNNs), for instance, are used to recognise faces, people, street signs, cancers, platypuses, and many other visual data elements

Project focuses on system gaining knowledge of set of rules to stumble on most cancers the usage of CT test photo. Detection of most cancers the usage of system gaining knowledge of strategies in CT test. Convolutional Neural Network(CNN) is a category of deep neural network, maximum generally implemented to investigate visible imagery. Its determined in literature survey that CNN, KNN and SVM offers 92%,84% and 88% of performance respectively in photo processing. CNN technique may be used to stumble on the lung most cancers.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, July 2022
Impact Factor: 6.252

II. LITERATURE SURVEY

Lung cancer detection using Machine Learning

Literature review plays a very vital role in the project. It mainly helps in gaining detailed knowledge about the basic ideas to focus on, and to collect information from the different perspective. By literature review we can get to know how to prioritise the work and complete it as intended. We can figure out the pros and cons of adopting a methodology and helps a lot in decision making and also making it more efficient. Conclusively literature review enables us to complete related literature review.

[1] A Review of most Recent Lung Cancer Detection Techniques using Machine Learning.
Nawzat Ahmed February 2021.International journal of Science and Business.
Lung most cancers is a form of risky most cancers and tough to detect. There were too many techniques advanced in latest years to diagnose lung most cancers, most of them utilizing CT scan images and some of them using x-ray images. In addition, multiple classifier methods are paired with numerous segmentation algorithms to use image recognition to identify lung cancer nodules. From the this study it has been found that CT scan images are more suitable to have accurate results. Therefore, mostly CT scan images are used for detection of cancer. the extracted features are fed to specified classifier to classify them as normal and malignant accordingly. Many classifiers have been used by the researchers in the literature such as: multi-layer perceptron (MLP), SVM, Naïve Bayes, Neural Network, Gradient Boosted Tree, Decision Tree, k-nearest neighbors, multinomial random forest classifier naïve Bayes, stochastic gradient descent, and ensemble classifier. it is clear that highest accuracy result was about 97% obtained by (Alam et al., 2018) using multi class SVM classifier as well as adopting marker-controlled watershed-based segmentation for image segmentation. On the other hand, all the works that have been implemented using Deep Learning methods obtained high accuracy results where the highest result was about 99% by (Li et al., 2020) using multi-resolution patch-based CNNs

[2] An Extensive Review on Lung Cancer Detection Using Machine Learning Techniques: A Systematic Study
Debnath Bhattacharya 24 March 2020.Revue d' Intelligence Artificielle
The Main Objective of this research paper is to investigate the accuracy levels of various machine learning algorithms. To find out the accuracy levels of various classifiers Based on the detection velocity for lung cancer using CT is 2.6 ten times greater than utilizing analogue radiography. To conquer the issues as well as to bring down the workload the methods recognized as the COMPUTER-AIDED DETECTION i.e .CAD methods are centered on the diagnosed information imaging progression as well as to sense the latent lesions in health.To test their model by using a collected 453 CT images of patients where 217 images were used as the training set the validation achieved a total accuracy of 82.9%.

[3] Lung Cancer detection using Machine Learning: Approach
Publisher: IEEE Smita Raut, Shraddha patil. January 2021. International journal of advance scientific research and engineering trends
In pre-processing, the input CT image is being processed to improve the quality of image. This enhanced version will contribute in further steps of any robotized system.
Image segmentation is the process in which a digital image is partitioned into multiple segments. In case of images segments corresponds to pixels or super pixels. In image processing, Otsu's method is used to automatically perform clustering-based image thresholding. It performs the reduction of a grey level image to a binary image. The algorithm works by assuming that there are two classes of pixels present in image following bi- modal histogram which includes foreground pixels and background pixels, it then computes the optimum threshold value which separates the two classes. Sobel filter is used for calculating gradient for edge detection. In IP special(Sobel) is used for sobel filtering. Grey-Level Co-Occurrence Matrix: A statistical mathematical method of examining feature texture that considers the spatial relationship of pixels in an image is the grey-level co- occurrence matrix (GLCM), also known as the grey-level spatial dependence matrix.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, July 2022
Impact Factor: 6.252

[4] Using Multi-level Convolutional Neural Network for Classification of Lung Nodules on CT images
Juan Lyu, Sai Ho Ling, Senior Member, IEEE 2018 IEEE.
Lung cancer is one of the four major cancers in the world. Accurate diagnosing of lung cancer in the early stage plays an important role to increase the survival rate. Computed Tomography (CT) is an effective method to help the doctor to detect the lung cancer. In this paper, we developed a multi-level convolutional neural network (ML-CNN) to investigate the problem of lung nodule malignancy classification. For ML-CNN, there are two convolution layers followed by batch normalization (BN) [21] and pooling layers. BN is used after the convolution operation and before the activation operation. It is used to reduce the internal covariate shift. The problem is formally known as covariate shift when the distribution of network activations changes between training and production stages.
In ML-CNN, there are 3 levels and they have same structures and same number of feature maps in the last convolution step. However, their convolutional kernels are different.

[5] Lung cancer Prediction and Classification based on Correlation Selection method Using Machine Learning Techniques
Qubahan academic journal Lung cancer is one of the leading causes of mortality in every country. This paper endeavors to inspect accuracy ratio of three classifiers which is Support Vector Machine (SVM), K- Nearest Neighbor (KNN)and, Convolutional Neural Network (CNN) that classify lung cancer in early stage so that many lives can be saving. The experimental results show that SVM gives the result 85.56%, CNN gives 92.11% and KNN gives 88.40%. The Confusion Matrix is a deep learning visual assessment method. The prediction class results are represented in the columns of a Confusion Matrix, whereas the real class results are represented in the rows . This matrix includes all the raw data regarding a classification model's assumptions on a specified data collection. To determine how accurate a model is. It's a square matrix with the rows representing the instances' real class and the columns representing their expected class. The confusion matrix is a 2 x 2 matrix that reports the number of true positives (TP), true negatives (TN), false positives (FP),and false negatives (FN) when dealing with a binary precision.

[6] Predicting Lung Cancer Survivability using SVM and Logistic Regression Algorithms. Avijith Mandal September 2017 International journal of computer applications

[7] Lung cancer Prediction and Classification Using Recurrent Neural Network. V Raaga Varsini, 11 November 2021.International journal of research in Engineering, Science and Management.
There are two types of lung cancer they are Small Cell Lung Cancer (SCLC) and Non-Small Cell Lung Cancer (NSCLC) this are the two main forms for lung cancer this will be develop and expand in their own ways. This non-small cell lung cancer has three subtypes they are (adenocarcinomas, squamous cell carcinomas, large cell carcinomas). The small/large cell cancer is a disease that occurs a patient and shows the symptoms for both types of cancer. (NSCLC) Adenocarcinoma is affect more common and it will be progressed more slowly than small cell lung cancer.

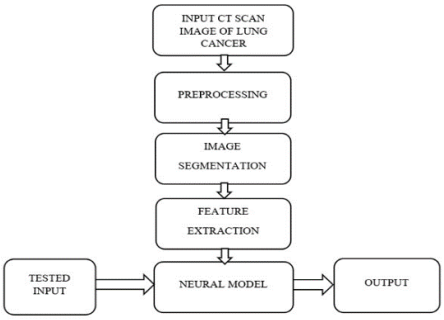
III. PROBLEM STATEMENT

Lung cancer begins in the lungs and may spread to lymph nodes or other organs in the body, such as the brain. Cancer from other organs also may spread to the lungs. When cancer cells spread from one organ to another, they are called metastases.
Lung cancers usually are grouped into two main types called small cell and non-small cell (including 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.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, July 2022
Impact Factor: 6.252

IV. METHODOLOGY

The system uses machine learning for its testing and training processes. The suggested model looks



The block diagram in figure 4.1 above is categorised as follows:

4.1 Pre-processing:

Pre-processing, as depicted in fig. 1, involves altering the raw CT image's quality. This involves doing specific procedures on the image in order to improve particular visual details and data.

4.2 Image Segmentation

The splitting of a digital image into separate portions is known as image segmentation. Photo segments are the same as pixels or super pixels. Segmentation is used to represent an image more simply, or in a form that is more significant and understandable.

When an image is segmented, it either produces a group of segments that together cover the entire image or a group of contours that are really drawn from the image. Regarding a characteristic or computed feature, such as colour, intensity, or texture, every pixel in a region is comparable. The hue of adjacent sections with relation to the same attribute differs dramatically. The contours produced after picture segmentation can be utilised with a stack of images, which is frequent in medical imaging, to produce 3D reconstructions using interpolation methods like marching cubes.

4.3 Feature Extraction

The process of turning raw data into processable numerical features while keeping the original data set's content intact. Compared to using machine learning on the raw data directly, it produces better outcomes.

4.4 Neural Model

Simple models of how the nervous system functions are called neural networks. A neural network is a streamlined representation of how the human brain functions. It simulates a huge number of connected processing units that mimic abstract representations of neurons in order to function.

Deep learning techniques are based on neural networks, sometimes referred to as artificial neural networks (ANNs) or simulated neural networks (SNNs), which are a subset of machine learning. Their structure and nomenclature are modelled after the human brain, mirroring the communication between organic neurons.

V. HARDWARE AND SOFTWARE REQUIREMENTS

Hardware Requirements

1. Laptop (core i3)
2. RAM 4 GB

Software Requirements

1. Google Collab
2. Python
3. Keras Library

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, July 2022
Impact Factor: 6.252

Colab: Google

As is obvious from the name, Google Colab is a Google product. Essentially, it is a cloud-based free notebook environment. It provides tools that make document editing similar to using Google Docs easy. Many well-known and advanced machine learning libraries are supported by Colab and are simple to load in your notebook.

Google Research's Colaboratory, also referred to as "Colab," is a product. Colab is particularly well suited to machine learning and data analysis because it enables anyone to develop and run arbitrary python code over the browser.

With Colab, we can fully utilise the capabilities of well-liked Python modules for data analysis and visualisation. We can import an image dataset into Colab and train an image.

Colab Applications:

1. Write and run Python code first.
2. Distribute the equation-supporting code you wrote.
3. Create, share, and upload notebooks.
4. Add notebooks to Google Drive and save them there.
5. Include external datasets.

Python

Python is an interpreted, object-oriented, high-level, dynamically semantic programming language. Python's straightforward syntax prioritises readability and makes it simple to learn, which lowers the cost of programme maintenance. Python's support for modules and packages promotes the modularity and reuse of code in programmes. Python is a popular computer programming language used to create software and websites, automate processes, and analyse data. Since Python is a general-purpose language, it may be used to develop a wide range of programmes and isn't tailored for any particular issues.

Kaggle

Kaggle Data Science Bowl (KDSB17) dataset is comprised of 2101 axial CT scans of patient chest cavities of the 2101, 1505 were initially released in stage 1 of the challenge, with 1397 belonging to the training set and 198 belonging to the testing set. The remaining 506 were released in stape 2 as a final testing set Each CT scan was labelled as with cancer' if the associated patient was diagnosed with cancer within I year of the scan, and without cancer otherwise. Crucially, the location or size of nodules are not labelled.

While the KDSB17 dataset provides CT scan image of patients, as well as their cancer status, it does not provide the locations or state of pulmonary nodules within the lung Therefore, in order to train our multi-stage framework, utilize an additional dataset, the Lung Nodule Analysis (LUNA16) dataset, This presents its own problems however, as this dataset does not contain the cancer status of patients.

Keras library

On top of open-source machine libraries like TensorFlow, Keras runs. The most well-known symbolic math package used to build neural networks and deep learning models is called TensorFlow. TensorFlow is incredibly adaptable, and distributed computing is its main advantage. It makes use of standalone machine learning toolkits, C#, Python, and C++ libraries. Although incredibly powerful, TensorFlow is a challenging library to use for building neural networks.

VI. CONCLUSION

In this study, we have shown a CNN neural network technique for early cancer detection. Due to its great accuracy and speed of computation, the CNN algorithm is well suited to decision-making for lung cancer screening. Video may be used to convey information more clearly than photos can. Early detection of lung cancer is advantageous since treatment may start right away to stop the disease from having negative effects. a review of the principal techniques for lung cancer prediction from CT imaging data and nodule classification..

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, July 2022
Impact Factor: 6.252

ACKNOWLEDGEMENT

We express our gratitude to our project guide Mrs. Pallavi J Assist, Professor, Department of Electronics and Communication Engineering, Vidya Vikas Institute of Technology & Engineering, who always stood behind us and supported us in every step of this project.

REFERENCES

[1]. Lung Cancer Symptoms, Types, Causes, Treatment & Diagnosis,
[2]. ICMR-National Institute of Cancer Prevention & Research (ICMRNICPR).
[3]. The International Agency for Research on Cancer (IARC). Latest global cancer data.
[4]. Top A.I Algorithms in Health Care-The Medical Futurists., 2019, https://medicalfuturist.com/top-ai-algorithms-healthcare/
[5]. G. Lakshmanaprabu S.K., Sachi Nandan Mohanty, Shankar K., Arunkumar N., Gustavo Ramirez., 2018. Optimal Deep Learning Model for Classification of Lung Cancer on CT Images. Future Generation Computer Systems,
[6]. Janee Alam, S., & Hossan, A. Multi-Stage Lung Cancer Detection and Prediction Using Multi-class SVM Classifier. 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2)
[7]. Janee Alam, S., & Hossan, A. Multi-Stage Lung Cancer Detection and Prediction Using Multi-class SVM Classifier. 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2)...
[8]. M.Gomathi, Dr.P.Thangaraj. 2010. A Computer-Aided Diagnosis System for Detection of Lung Cancer Nodules using Extreme Learning Machine. International Journal of Engineering Science and Technology. Vol. 2(10), 2010, 5770- 5779.
[9]. Kakeda, S., Moriya, J., Sato, H., Aoki, T., Watanabe, H., Nakata, Doi, K. 2004. Improved Detection of Lung Nodules on Chest Radiographs Using a Commercial Computer-Aided Diagnosis System. American Journal of Roentgen logy, 182(2), 505– 510. doi:10.2214/ajr.182.2.1820505.
[10]. Gurcan, M. N., Sahiner, B., Petrick, N., Chan, H.-P., Kazerooni, E. A., Cascade, P. N., & Hadjiiski, L. 2002. Lung nodule detection on thoracic computed tomography images: Preliminary evaluation of a computeraided diagnosis system. Medical Physics, 29(11), 2552– 2558. doi:10.1118/1.1515762.