LUNG CANCER DETECTION USING MACHINE LEARNING

A PROJECT REPORT

Submitted by

P. MUTHUPRIYA – 2019202035

in partial fulfilment or the award of the degree

of

MASTER OF COMPUTER APPLICATION

DEPARTMENT OF INFORMATION SCIENCE AND TECHNOLOGY,

COLLEGE OF ENGINEERING, GUINDY

ANNA UNIVERSITY, CHENNAI 600 025.

May,2022.

ANNA UNIVERSITY CHENNAI - 600 025

BONAFIDE CERTIFICATE

Certified that this project report titled LUNG CANCER DETECTION USING MACHINE LEARNING is the bonafide work of MUTHUPRIYA who carried out project work under my supervision. Certified further that to the best of my knowledge and belief, the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or an award was conferred on an earlier occasion on this or any other candidate.

Place: Chennai

Date:30-05-2022

DR L SAI RAMESH

TEACHING FELLOW

PROJECT GUIDE DEPARTMENT OF IST,

CEG ANNA UNIVERSITY CHENNAI 600025

COUNTER SIGNED

Dr. S. SRIDHAR

HEAD OF THE DEPARTMENT

DEPARTMENT OF INFORMATION SCIENCE AND TECHNOLOGY

COLLEGE OF ENGINEERING, GUINDY ANNA UNIVERSITY

CHENNAI 600025

TABLE OF CONTENTS

ABSTRACT

Lung cancer is one of the dangerous and life taking disease in the world. However, early diagnosis and treatment can save life. Although, CT scan imaging is best imaging technique in medical field, it is difficult for doctors to interpret and identify the cancer from CT scan images. Therefore, computer aided diagnosis can be helpful for doctors to identify the cancerous cells accurately.

The proposed system works using images processing techniques, SVM classifier and Decision tree algorithm. Image processing techniques used to improve the quality of image and make the image in the better format for feature extraction. In feature extraction we show the Extracted tumour, Boundary detection, Circularity of the tumour, Diameter of the tumour & Stage of the tumour. SVM classifier is the binary classification technique used to detect the benign and malignant tumour. If the tumour is malignant, it was further classified into which type of lung cancer using Convolutional Neural Network. Using diameter of the tumour the stage of cancer can also be determined.

LIST OF FIGURES

Chapter 1

INTRODUCTION

1.1 General:

Machine learning (ML) allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so. Machine learning algorithms use historical data as input to predict new output values. Machine learning has several practical applications that drive the kind of real business results - such as time and money savings - that have the potential to dramatically impact the future of your organization.

Healthcare: Machine Learning is widely used in the healthcare industry. It helps healthcare researchers to analyze data points and suggest outcomes.

Automation: This is one of the significant applications of machine learning that helps to make the system automated. It helps machines to perform repetitive tasks without human intervention.

Cancer is a type of disease which destroys cells in our body. There are various types of cells, which will cause different types of cancers in our body. An abnormal growth in the cells will cause cancer and it will grow rapidly. A set of the cancer will form a tumour which will affect the normal health tissues. There are two types of tumour cells which are classified as benign or malignant. The malignant tumours refer to the cancer where benign tumours will not affect the other cells. There are two types of lung cancer: small cell lung cancers (SCLC) and non-small lung cancer (NSCLC). It is classified based on the size of the cell in microscopic appearance. The stages of lung cancer can be classified based on how far it is spread. Thus, early detection will improve the survival rate of the cancerous patients.

1.2 Problem Statement:

CT scan imaging is best imaging technique in medical field, it is difficult for doctors to interpret and identify the cancer from CT scan images. Therefore, computer aided diagnosis can be helpful for doctors to identify the cancerous cells accurately. CAD will

eliminate the distortion and unwanted noise in the image using image processing techniques. Then the classifier is used to classify whether the tumour is benign or malignant and its type.

1.3 Objective:

To design and develop a system to classify whether the cancerous cells present in the given CT scan or not and its type using Machine Learning algorithms.

Chapter 2

LITERATURE REVIEW

2.1 Image Acquisition:

Image Acquisition [1] The first stage of methodology starts with collecting the datasets (CT images). The datasets contain normal and abnormal images. The images are in raw data format. So, it needs to pre-processed the images to improve contrast transparency

2.2 Image Pre-Processing:

Pre-processing refers to the transformations applied to our data before feeding it to the algorithm. Data pre-processing [1] is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis. Smoothing is an image processing technique used in order to reduce noise in an image to produce clearer image. Most of the techniques are based on low pass linear filters. Median filter [2] is used for smoothing, to remove noise and unwanted distortion in the image. Gabor filter [3] named by Gabor, may be a linear filter is employed for edge detection and texture analysis. Representation of Gabor filter almost like the human sensory system.

Segmentation [5] is a process which splits the image into its constituent regions or objects. Segmentation is usually used to trace objects and borders such as lines, curves, etc. in images. The main objective of segmentation is to simplify and change the representation of the image into something that is more significant and easier to examine. In our proposed system, Segmentation involves two parts which are Lung Segmentation and Nodules Segmentation. Lung Segmentation [4] is implemented through HU Scale where once the image is converted into HU Scale, we get segmented binary image of lungs. Further in Nodules Segmentation, nodules are detected within the binary image by using Morphological Operations such as dilation and erosion. Dilation and Erosion are often used in combination for specific image pre-processing applications such as filing holes or removing small objects. As a result of which, we get segmented nodule image given to feature extraction to perform textual analysis on the image.

2.3 Feature Extraction:

Feature Extraction [6] is a method by which we aim at reducing the number of dimensions that our raw data contains so that it is easier to process and is in a form of manageable classes. It uses different methods and algorithms for feature extraction from the segmented image. The extracted image can be classified as either cancerous or noncancerous using texture properties. Binarization [5] is the process of changing the colour of the pixel values into two classes such as black and white. Binarization converts the image to grey scale image and the threshold value is determined which is compared with threshold value of normal lung to conclude the tumour present or not.

2.4 Classification:

This stage classifies the detected nodule as malignant or benign. Support Vector Machine (SVM) is used as a classifier. SVM [5] is supervised machine learning algorithm which defines the function that classifies data into two classes. In our proposed system, we have defined two classes as cancerous or non-cancerous. SVM is a binary classification method that takes as input labelled data from two classes and outputs a model file for classifying unknown or known data into one of two classes.

If the SVM result is cancerous then the type of cancer is determined using CNN. Convolutional Neural Networks [6] come under the subdomain of Machine Learning which is Deep Learning. Algorithms under Deep Learning process information the same way the human brain does, but obviously on a very small scale, since our brain is too complex. A neuron is the basic element of a neural network. A neuron takes in input, applies weight to it to predict the output. Each node in the hidden layer is a function of the output from the preceding layer, in this case, the input layer. The function applied to obtain the output at the hidden layers is called the activation function. Finally, the output $\boldsymbol{y}$ is obtained by applying weights to each node at the hidden layer and combining them by applying an appropriate activation function.

Chapter 3

OVERALL ARCHITECTURE

3.1 Architecture Design:

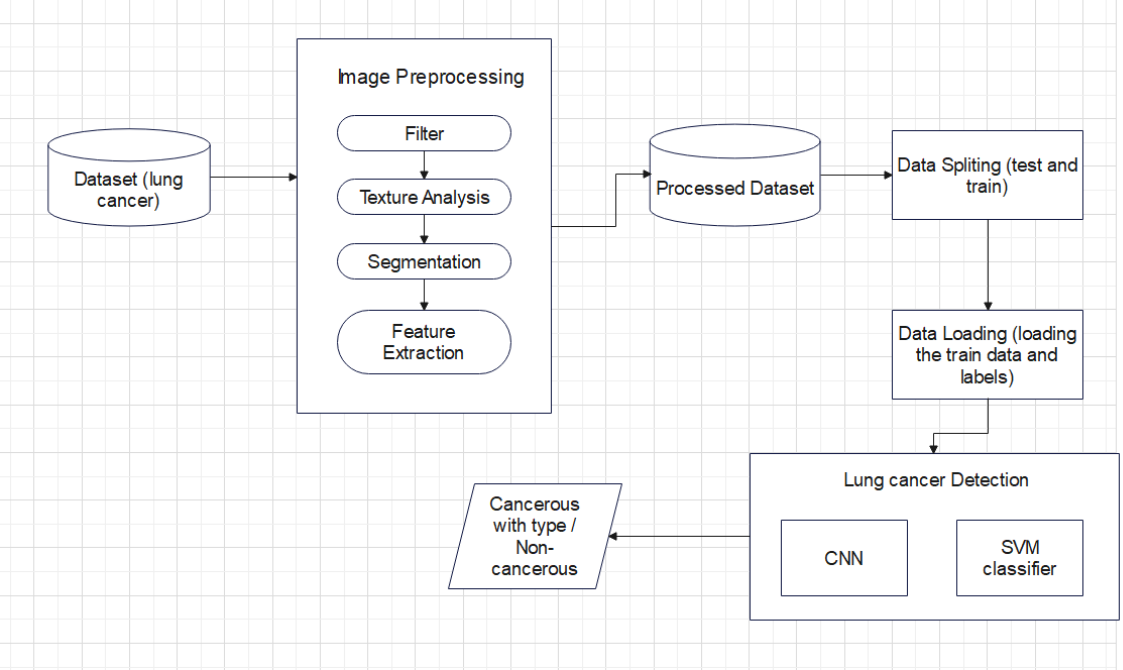

Fig 3.1, Architecture Diagram

3.2 Architecture Explanation:

The Diagram Fig 3.1 shows the Architecture of the proposed System. Initial step of the proposed system is to pre-process the CT scan images. The image Pre-processing includes filtering, Texture analysis, segmentation and nodule detection. Median filter is used to smoothing the image and suppress unwanted noise. Gabor filter is used for texture analysis. Segmentation is done through region growing and watershed algorithm. Binarization is used for feature extraction. Thus, the processed data is split into test and train datasets. The data loader loads these train and test datasets to the SVM classifier. If the tumour is malignant further detect the diameter of the tumour. Based on that the types of lung cancer is determined. CNN is used to determine the type of lung cancer. The Model evaluation is done by confusion matrix and accuracy is determined.

3.3 List of Modules:

- ➢ Data Acquisition
- ➢ Pre-processing
- ➢ Feature extraction
- ➢ Classification
- ➢ Prediction

3.4 Modules Explanation:

Data Acquisition:

- CT image is preferred for lung cancer detection rather than other available medical images.
- There are many publicly available databases build to help the researcher. They can utilize those datasets for training and testing of their algorithms and models. Among those databases, most commonly used are LIDC, TCIA, and Kaggle.
- Using os python library the data is appended into the system for further image processing and classification.

Pre-Processing:

- The Pre-processing makes the data more reliable to use it. Some filters like median filter are used to clean, smooth, eliminate noise.
- Gabor filter is used for texture analysis. watershed algorithm is used for segmentation.

Feature Extraction:

- Binarization is the process of changing the color of the pixel values into two classes such as black and white. Binarization converts the image to grey scale image and the threshold value is determined which is compared with threshold value of normal lung to conclude the tumour present or not. Its working is shown in Fig 3.2.

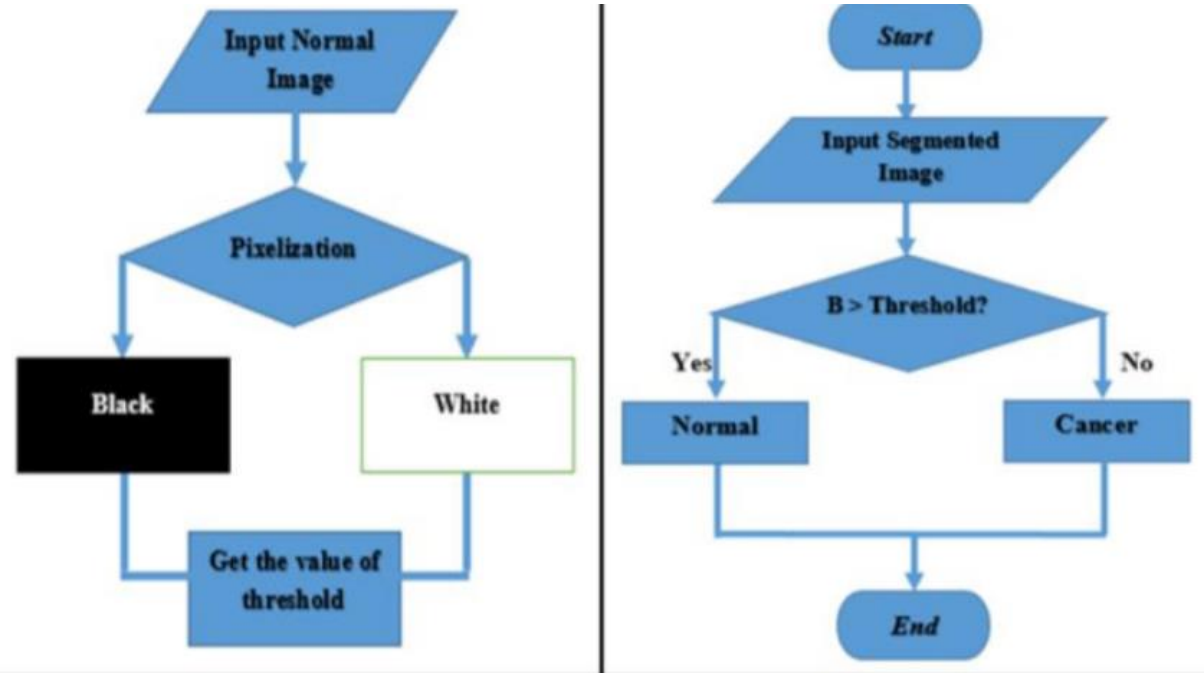

Fig 3.2, Binarization

Classification:

- SVM is a supervised machine learning algorithm which can be used for classification or regression problems. Here, the processed data undergoes train and test.

Chapter 4

IMPLEMENTATION

4.1 Framework / Platform

- Language: Python
- Jupiter Notebook

4.2 Modules Algorithm

SVM:

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the Fig 4.1 in which there are two different categories that are classified using a decision boundary or hyperplane:

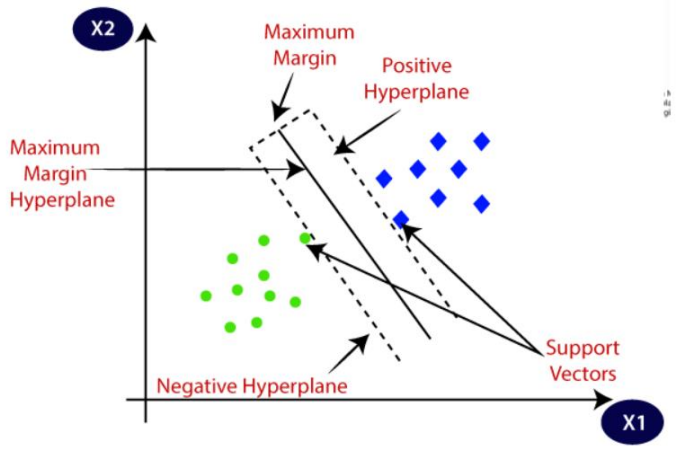

Fig 4.1, SVM Classifier

- Steps in SVM implementation

Step1: start.

Step2: Read the dataset.

Step3: Label the data if necessary.

Step4: Split the data for train (75%) and test (25%).

Step5: Train the data.

Step6: After trained the data successfully, Test the data.

Step7: Check accuracy.

Step8: End.

CNN:

Convolutional Neural Networks is the subdomain of Machine Learning which is Deep Learning. It can take an input image, assign importance (learnable weights and biases) to various aspects/objects in the image and be able to differentiate one from the other.

Architecture of CNN:

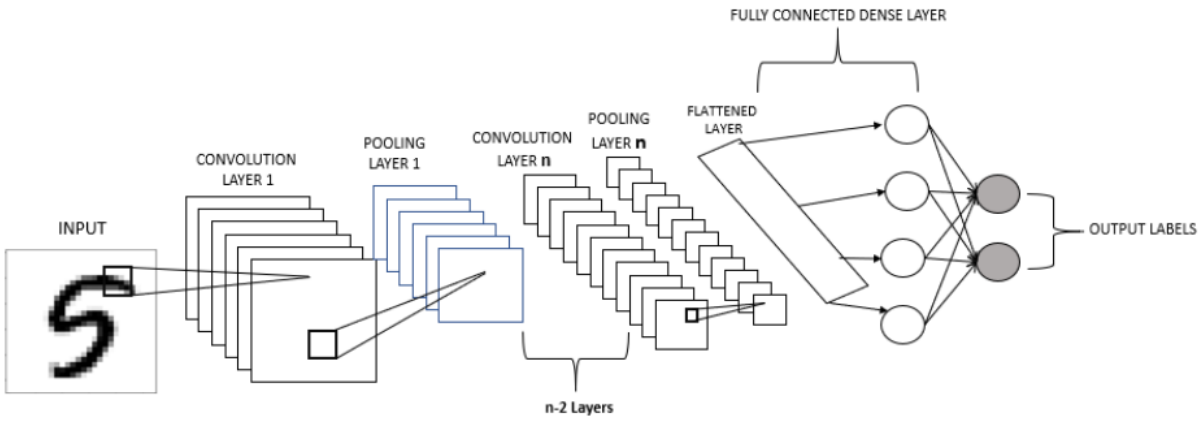

Fig 4.2, Architecture of CNN

The convolution layers extract features from an image (input). A small filter or kernel scans through the image and extracts features, for example, a vertical or a horizontal line, and creates a feature map. The layer that comes after the convolution layer is the pooling layer. The pooling layer essentially down samples the feature map extracted by the convolution layer. A feature map extracted by the convolution layer contains the exact position of the feature. This may result in overfitting. The pooling layer runs a filter across the features map. It only takes the specific information from that filter, either the average of the values coming under the filter or the maximum depending on the pooling method selected. This reduces the feature map's spatial size and translates the feature's exact spatial information to rough information. This helps to prevent overfitting. Any convolution and pooling layers can be stacked together depending on the complexity of the input data set. The final pooling layer is then flattened out and transformed into a one-dimensional array, and fed to the fully connected layers that predict the output.

4.3 SCREENSHOTS

The image in Fig 5.1 shows the pre-processed image.

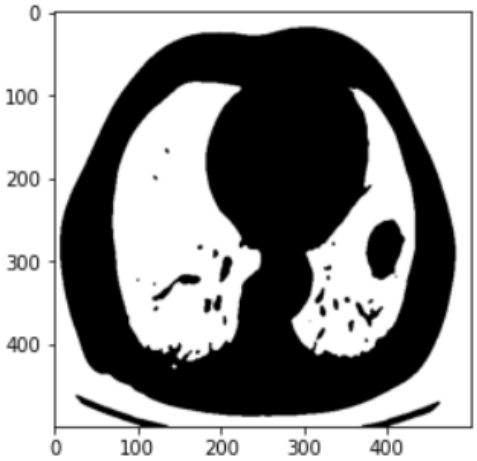

Fig 5.1, pre-processed image

SVM Result:

```
In [26]: #Accuracy and prediction

In [27]: accuracy = model.score(xtest,ytest)

         print('Accuracy:',accuracy)

         print('Prediction is : ',categories[prediction[0]])

Accuracy: 0.9920318725099602
Prediction is :  Cancerous
```

CNN Result:

```
In [156]: score = model.evaluate(X_test, y_test, verbose = 0 )
          print("Test Score: ", score[0])
          print("Test accuracy: ", score[1])

Test Score:  0.48521071672439575
Test accuracy:  0.8552631735801697
```

```
In [158]: print("Prediction: ",CATEGORIES[pred[1]])

Prediction:  adenocarcinoma
```

Chapter 5

REFERENCES

[1] Dr. M. Sangeetha, Department of Information Technology “Classification of lung cancer” in International Journal of Engineering Research & Technology,2020.

[2] Pratyaksh Jain, “Lungs Cancer Detection System” in IRJET,2020.

[3] F. Taher, N. Prakash, A. Shaffie, A. Soliman, A. El-Baz, “An Overview of Lung Cancer Classification Algorithms and their Performances” in IAENG International Journal of Computer Science,2021.

[4] Prerana Prajapati, Vedika Hande, Aarti Ingale, Sanjeev Dwivedi, “Lung Cancer Detection and Classification Using SVM” in JETIR,2019.

[5] Pragya Chaturvedi, “Prediction and Classification of Lung Cancer Using Machine Learning Techniques” in IOP Conference Series: Materials Science and Engineering,2021.

[6] Vipul Kumar,” Image Classifier with Convolutional Neural Network (CNN)” in Towards Data Science,2021.