# ABSTRACT

**Lung cancer detection at right time is a highly difficult task as it shows no such symptoms in the beginning.** Detection of lung cancer at an early stage can be helpful to save many lives. Computed Tomography (CT) scanned images are one of the most preferable to the radiologist to diagnose any abnormalities on the lung. Doctors find it difficult to detect abnormalities in CT scanned images manually and it can be erroneous. **Therefore, Computer-Aided Techniques are required for the proper detection of lung diseases. In this paper, we propose a novel approach to the detection and prediction of lung cancer by image processing of the CT scanned images to a certain level. Different pre-processing techniques have been applied to smoothen and enhance the images. We applied this processed image dataset into different deep learning models and machine learning models to classify them into three groups (Normal, Benign, and Malignant). Here** we have also done vgg16 feature extract with **different deep learning and machine learning models achieve highest accuracy rate. After comparing we find remarkable accuracy in recognition of lung tumor and prediction of the severity level using SVM and logistic regression model.**

**Keywords:** Computed Tomography (CT), **SVM, logistic Regression, Vgg-16**

# ACKNOWLEDGEMENT

This diligence wouldn’t have been doable without a number of people whose collaboration is received. Initially of all, we wish to thank the supreme power of the Almighty Allah who is clearly the one who at all times guide us to exertion on the correct path of life. Without his grace this research couldn’t become reality. We are feeling obliged in taking the opportunity to sincerely gratitude our supervisor Dr. Umma Hany, Associate Professor, Department of Electrical & Electronics Engineering, Ahsanullah University of Science & Technology (AUST) for her helpful contributions, valuable guidance, suggestions and continuous interest throughout this research. We would like to thank the Cancer Imaging Archive (TCIA) & Kaggle dataset site for the accessible CT scan image database of Lung disease patients. Our cheers and greatest wishes are additionally conveyed to our classmates and wonderful friends, seniors for sharing their information and inspiring us. Lastly, yet importantly, we wish to particular deep gratitude to our beloved mother and father for a life-long love and affection. We’d additionally wish to thank for his or her without stopping inspiration and support, relating to the completion of this thesis all over the years of our studies.

Dhaka

April, 2021

Thankfully

Nusraat Nawreen

Tahmina Islam

Nafee Muhtasim

Abdullah Al Mamun

# 

# Table of Contents

[ABSTRACT i](#_Toc70172742)

[ACKNOWLEDGEMENT ii](#_Toc70172743)

[Table of Contents iii](#_Toc70172744)

[List of Figures vi](#_Toc70172745)

[List of Tables ix](#_Toc70172746)

[Abbreviations and Acronyms xi](#_Toc70172747)

[Chapter 1 1](#_Toc70172748)

[Introduction 1](#_Toc70172749)

[1.1 Background of the Research 1](#_Toc70172750)

[1.2 Objective of the Research 3](#_Toc70172751)

[1.3 Scope and limitation of the Research 3](#_Toc70172752)

[1.4 Study Area Profile 3](#_Toc70172753)

[1.5 Structure of the Thesis 4](#_Toc70172754)

[Chapter 2 6](#_Toc70172755)

[Theoretical Framework and Methodology 6](#_Toc70172756)

[2.1 Basic Terminologies 6](#_Toc70172757)

[2.1.1 Lung Cancer 6](#_Toc70172758)

[2.1.2 Lung Tumor Classification 7](#_Toc70172759)

[2.2 Literature Review 7](#_Toc70172760)

[2.3 Methodology of the Research 9](#_Toc70172761)

[2.4 Detection of nodule using MATLAB 10](#_Toc70172762)

[2.4.1 Proposed methodology 10](#_Toc70172763)

[2.4.2 Methodologies 11](#_Toc70172764)

[A. Image Pre-Processing 11](#_Toc70172765)

[ Image Resizing: 11](#_Toc70172766)

[ Image Smoothing 11](#_Toc70172767)

[ Image Enhancement: 11](#_Toc70172768)

[B. Image Segmentation 12](#_Toc70172769)

[ Thresholding: 12](#_Toc70172770)

[ Edge Detection: 13](#_Toc70172771)

[D. Classification 13](#_Toc70172772)

[2.5 Classification of lung tumor using deep learning and machine learning algorithms: 14](#_Toc70172773)

[2.5.1 Image Processing: 14](#_Toc70172774)

[ Image Resizing: 14](#_Toc70172775)

[ Median Filtering: 14](#_Toc70172776)

[ Image Segmentation: 15](#_Toc70172777)

[Dataset Visualization 16](#_Toc70172778)

[2.6 Machine learning: 18](#_Toc70172779)

[2.6.1 CNN 18](#_Toc70172780)

[2.6.2 Vgg16: 19](#_Toc70172781)

[2.6.3 K-Nearest Neighbors Classifier 20](#_Toc70172782)

[2.6.4 SVM (support vector machine) 21](#_Toc70172783)

[2.6.5 Decision Tree 21](#_Toc70172784)

[2.6.6 Naive Bayes 22](#_Toc70172785)

[2.6.7 Logistic Regression 23](#_Toc70172786)

[2.7 System Architecture 24](#_Toc70172787)

[CNN with Processed Dataset 27](#_Toc70172788)

[CNN without Processed Dataset 28](#_Toc70172789)

[Support Vector Machine on processed Dataset 30](#_Toc70172790)

[Logistic Regression Classifier on processed Dataset 31](#_Toc70172791)

[Random Forest Classifier on processed Dataset 32](#_Toc70172792)

[Unprocessed Dataset 33](#_Toc70172793)

[Support Vector Machine on Unprocessed Dataset 33](#_Toc70172794)

[Logistic Regression on unprocessed Dataset 34](#_Toc70172795)

[Random Forest Classifier on unprocessed Dataset 35](#_Toc70172796)

[Vgg16 feature extraction with Processed Dataset 36](#_Toc70172797)

[Support Vector Machine on Vgg16 feature extraction with Processed Dataset 36](#_Toc70172798)

[Random Forest Classifier on Vgg16 feature extraction with Processed Dataset 37](#_Toc70172799)

[Logistic Regression on Vgg16 feature extraction with Processed Dataset 38](#_Toc70172800)

[Support Vector Machine on Vgg16 feature extraction without Processed Dataset 39](#_Toc70172801)

[Logistic Regression on Vgg16 feature extraction without Processed Dataset 40](#_Toc70172802)

[Chapter 3 43](#_Toc70172803)

[Monitoring and Analysis of system model 43](#_Toc70172804)

[3.1 Feature Extraction from Image processing 43](#_Toc70172805)

[3.2 Monitoring from image processing 44](#_Toc70172806)

[3.3 Analysis of SVM from Matlab tools 47](#_Toc70172807)

[3.4 Classification Report: 48](#_Toc70172808)

[Chapter 4 52](#_Toc70172809)

[Results and Discussion 52](#_Toc70172810)

[4.1 Matlab Result 52](#_Toc70172811)

[4.2 Result of proposed model: 53](#_Toc70172812)

[Deep learning (CNN) Classification Results 68](#_Toc70172813)

[4.3 Classification Report with chart 68](#_Toc70172814)

[Chapter 5 72](#_Toc70172815)

[Conclusion and Recommendation 72](#_Toc70172816)

[5.1 CONCLUSION 72](#_Toc70172817)

[5.2 Limitations 73](#_Toc70172818)

[5.3 Recommendations 73](#_Toc70172819)

[References 74](#_Toc70172820)

[References 74](#_Toc70172821)

# List of Figures

[Figure 2.1: Lung Cancer Types 7](#_Toc70169778)

[Figure 2.2: Propossed system model 10](file:///C:\Users\USER\Downloads\finallllllll.docx#_Toc70169779)

[Figure 2.3: SVM Classifier 13](#_Toc70169780)

[Figure 2.4: Example of Image Segmentation 15](#_Toc70169781)

[Figure 2.5: Example of Image Segmentation 16](#_Toc70169782)

[Figure 2.6: Dataset Visualization 17](#_Toc70169783)

[Figure 2.7: PI Chart of Data Visualization 17](#_Toc70169784)

[Figure 2.8: Convolutional Neural Network Model 19](file:///C:\Users\USER\Downloads\finallllllll.docx#_Toc70169785)

[Figure 2.9: Network Structure 20](#_Toc70169786)

[Figure 2.10: Support Vector Machine 21](#_Toc70169787)

[Figure 2.11: General view of decision tree 22](#_Toc70169788)

[Figure 2.12: Example X data and Probability-Class Relations 23](#_Toc70169789)

[Figure 2.13: Logistic Regression to distinguish two classes 24](#_Toc70169790)

[Figure 2.14: System Architecture 25](#_Toc70169791)

[Figure 15 Overview of this study 25](file:///C:\Users\USER\Downloads\finallllllll.docx#_Toc70169792)

[Figure 2.16: Example of Processed Image of the Test Dataset Using CNN with Processed Dataset 27](#_Toc70169793)

[Figure 2.17: Example of Mispredicted Images of the Test Data using CNN with Processed Dataset 28](#_Toc70169794)

[Figure 2.18: Examples of Predicted Image of the Test Dataset using CNN without Processed Dataset 28](#_Toc70169795)

[Figure 2.19: Example of Mispredicted Images of the Test Data using CNN without Processed Dataset 29](#_Toc70169796)

[Figure 2.20: Example of predicted Images of the Test Data using SVM ont Processed Dataset 30](#_Toc70169797)

[Figure 2.21: Example of Mispredicted Images of the Test Data using CNN without Processed Dataset 30](#_Toc70169798)

[Figure 2.22: Example of predicted Images of the Test Data using Logistic Regression Classifier on Processed Dataset 31](#_Toc70169799)

[Figure 2.23: Example of Mispredicted Images of the Test Data using Logistic Regression Classifier on Processed Dataset 31](#_Toc70169800)

[Figure 2.24: Example of predicted Images of the Test Data using Random Forest Classifier on processed Dataset 32](#_Toc70169801)

[Figure 2.25: Example of Mispredicted Images of the Test Data using Random Forest Classifier on processed Dataset 32](#_Toc70169802)

[Figure 2.26: Example of predicted Images of the Test Data using Support Vector Machine on Unprocessed Dataset 33](#_Toc70169803)

[Figure 2.27: Example of Mispredicted Images of the Test Data using Support Vector Machine on Unprocessed Dataset 33](#_Toc70169804)

[Figure 2.28: Example of predicted Images of the Test Data using Logistic Regression on unprocessed Dataset 34](#_Toc70169805)

[Figure 2.29: Example of Mispredicted Images of the Test Data using Logistic Regression on unprocessed Dataset 34](#_Toc70169806)

[Figure 2.30: Example of predicted Images of the Test Data using Random Forest Classifier on unprocessed Dataset 35](#_Toc70169807)

[Figure 2.31: Example of Mispredicted Images of the Test Data using Random Forest Classifier on unprocessed Dataset 35](#_Toc70169808)

[Figure 2.32: Example of predicted Images of the Test Data using Support Vector Machine on Vgg16 feature extraction with Processed Dataset 36](#_Toc70169809)

[Figure 2.33: Example of MIspredicted Images of the Test Data using Support Vector Machine on Vgg16 feature extraction with Processed Dataset 36](#_Toc70169810)

[Figure 2.34: Example of predicted Images of the Test Data using Random Forest Classifier on Vgg16 feature extraction with Processed Dataset 37](#_Toc70169811)

[Figure 2.35: Example of MIspredicted Images of the Test Data using Random Forest Classifier on Vgg16 feature extraction with Processed Dataset 37](#_Toc70169812)

[Figure 2.36: Example of predicted Images of the Test Data using Logistic Regression on Vgg16 feature extraction with Processed Dataset 38](#_Toc70169813)

[Figure 2.37: Example of Mispredicted Images of the Test Data using Logistic Regression on Vgg16 feature extraction with Processed Dataset 38](#_Toc70169814)

[Figure 2.38: Example of predicted Images of the Test Data using Support Vector Machine on Vgg16 feature extraction without Processed Dataset 39](#_Toc70169815)

[Figure 2.39: Example of Mispredicted Images of the Test Data using Support Vector Machine on Vgg16 feature extraction without Processed Dataset 39](#_Toc70169816)

[Figure 2.40: Example of predicted Images of the Test Data using Logistic Regression on Vgg16 feature extraction without Processed Dataset 40](#_Toc70169817)

[Figure 2.41: Example of Mispredicted Images of the Test Data using Logistic Regression on Vgg16 feature extraction without Processed Dataset 40](#_Toc70169818)

[Figure 2.42: Example of predicted Images of the Test Data using Random Forest Classifier on Vgg16 feature extraction without Processed Dataset 41](#_Toc70169819)

[Figure 2.43: Example of Mispredicted Images of the Test Data using Random Forest Classifier on Vgg16 feature extraction without Processed Dataset 41](#_Toc70169820)

[Figure 3.44: Two of the Lung disease CT images and their Histogram 45](#_Toc70169821)

[Figure 3.45: Cropped image and the median filtered image of the CT images 45](#_Toc70169822)

[Figure 3.46: The Enhanced Images and their Histograms 46](#_Toc70169823)

[Figure 3.47: Edges of the Binary images and the Edge subtracted Binary images 47](#_Toc70169824)

[Figure 3.48: ROI of the Lung Tumor and their Classification 48](#_Toc70169825)

[Figure 3.49: Accuracy Score comparison with Machine Learning approach 51](#_Toc70169826)

[Figure 4.50: Processed Dataset- Benign case classification results 54](#_Toc70169827)

[Figure 4.51: Without Processed Dataset- benign case classification results 54](#_Toc70169828)

[Figure 4.52: Vgg16 Feature Extraction with processed dataset-benign case classification results 55](#_Toc70169829)

[Figure 4.53: Vgg16 Feature Extraction without Processed Dataset-benign case classification results 55](#_Toc70169830)

[Figure 4.54: Machine Learning with processed dataset-malignant case detection result 57](#_Toc70169831)

[Figure 4.55: Machine Learning with Processed Dataset-Malignant Case Detection Result 57](#_Toc70169832)

[Figure 4.56: Vgg16 Feature Extraction with Processed Dataset-Malignant Case Detection Result 58](#_Toc70169833)

[Figure 4.57: Vgg16 Feature Extraction without Processed Dataset-Malignant Case Detection Result 58](#_Toc70169834)

[Figure 4.58: Processed Dataset-Normal Case Classification Result 60](#_Toc70169835)

[Figure 4.59: Without Processed Dataset-Normal Case Classification Result 60](#_Toc70169836)

[Figure 4.60: Vgg16 Feature Extraction with Processed Dataset-Normal Case Classification Result 61](#_Toc70169837)

[Figure 4.61: Vgg16 Feature Extraction without Processed Dataset-Normal Case Classification Result 61](#_Toc70169838)

[Figure 4.62: Deep learning (CNN) Approach Accuracy Graph 63](#_Toc70169839)

[Figure 4.63: Plot of Model Accuracy vs. Epoch for Training and Validation Images 64](#_Toc70169840)

[Figure 4.64: Plot of Model Loss vs. Epoch for Training and Validation Images 64](#_Toc70169841)

[Figure 4.65: Accuracy Curve for CNN model on without Processed Dataset 65](#_Toc70169842)

[Figure 4.66: Loss for CNN model on without Processed Dataset 65](#_Toc70169843)

[Figure 4.67: Accuracy Curve for VGG16 Feature extraction with processed Dataset 66](#_Toc70169844)

[Figure 4.68: Loss Curve for VGG16 Feature extraction with processed Dataset 66](#_Toc70169845)

[Figure 4.69: Accuracy Curve for VGG16 Feature extraction without processed Dataset 67](#_Toc70169846)

[Figure 4.70: Loss Curve for VGG16 Feature extraction without processed Dataset 67](#_Toc70169847)

[Figure 4.71: CNN with Processed dataset 69](#_Toc70169848)

[Figure 4.72: CNN without Processed Dataset 70](#_Toc70169849)

[Figure 4.73: Vgg16 Feature Extraction with Processed Dataset 71](#_Toc70169850)

[Figure 4.74: Vgg16 Feature Extraction without Processed Dataset 71](#_Toc70169851)

# List of Tables

[Table 1: Classfication based on training features and the extracted test features 48](#_Toc70166389)

[Table 2: Accuracy Score comparison with Machine Learning approach 50](#_Toc70166390)

[Table 3: Nodules classification by proposed method 53](#_Toc70166391)

[Table 4: Benign case classification results 53](#_Toc70166392)

[Table 5: Malignant case classification results 56](#_Toc70166393)

[Table 6: Normal case classification results 59](#_Toc70166394)

[Table 7: Deep learning (CNN) approach accuracy table 62](#_Toc70166395)

[Table 8: Deep learning (CNN) classification results 68](#_Toc70166396)

# Abbreviations and Acronyms

**CNN** Convolutional Neural Network

**CT** Computed Tomography

**ANN** Artificial Neural Network

**SCLC** Small cell lung cancer

**NSCLC** Non-small cell lung cancer

**SVM** Support Vector Machine

**MLP** Multi-Layer Perceptron

**ROI** Region of Interest

**DT** Decision Tree

**KNN** K-nearest Neighbor

**FC** Fully-Connected

**GLM** General Linear Model regression

**BPNN** Back Propagation Neural Network

**FPSO**  Fuzzy particle swarm optimization

**LR** logistic regression

**AI** artificial intelligence

**BN** bunch standardization

**NB** Naïve Bayes

**LR** Logistic Regression

## Chapter 1

### Introduction

#### 1.1 Background of the Research

Lung cancer was first found in 1761 [1] . WHO reported that in 2018 1,368,524cases were found in men which covers 14.5% of total cancer cause for men and for women 725,352cases were found which covers 8.4% of total cancer cause for women [2]. In 2020 there were 2,206,771 (11.4%) lung cancer cases found globally [3]. Lung cancer needs to be detected as early as possible [4]. Lung cancer can be detected by using SVM classifier [5]. Automatic convolutional neural network (CNN) [6] based operation like AlexNet [7], VGG [8], GoogleNet [9] performed medical image analysis tasks successfully [10]. There are also 3D deep encoder-decoder CNN architecture which can use CAD system properly in clinical application [11]. By doing automatic interpretation of chest radiograms abnormalities can be found in lung [12]. Flat-types of the electrodes can also estimate cancer tissue [13].

The cost of lung cancer care is a big issue for patient and their family. There are many clinical phases of lung cancer like pre-diagnosis, staging, initial, continuing, and terminal. The Surveillance, Epidemiology, and End Results (SEER)-Medicare data which was collected from 1991 to 2003 for 60,231 patients having lung cancer were used to estimate the costs for clinical phases [14]. The results showed that monthly cost for a 72 year old patient ranged from $2687 (no treatment) to $9360 (chemo-radiotherapy). In 2016 the total cost for a lung cancer patient was $210,067. A patient has to pay for his/ her primary care doctor, a pulmonologist, a medical oncologist, a palliative care specialist, and the doctors in the emergency room. The primary approaches for treating cancer is surgery, radiation, and pharmacological therapy like chemotherapy, targeted therapy, hormone therapy and immunotherapy. Cost depends on various types of therapy patient taking on. Some patient need one or two therapy, some may need all of the therapy existed.

Machine learning takes AI software a step further as it enables intelligent learning to occur within the component based on previous work it did or extrapolations made from data. The software performs sophisticated decision making processes as it goes along and learns from previous activities. A brief description of the research papers based on Lung Cancer detection using different Machine learning algorithms are explained below:

Deals with the prediction of post-operative life expectancy in lung cancer patients using predictive data mining algorithms to compare algorithms such as Decision Tree, Naive Bayes and Artificial neural network. A stratified 10-fold cross-validation comparative analysis was conducted on the above algorithms and accuracy was calculated for each classifier [15].

Paper deals with comparative study of classification algorithm for detection of Brain Tumor. Using volumetric and location features overall accuracy rate was calculated based on 2 classification classes such as logistic regression and Quadratic Discriminant and 3 classification classes such as Linear SVM, Coarse Gaussian SVM, Cosine KNN and Complex and median tree [16].

In this paper, different results are produced for each classifier on the lung cancer dataset obtained. The classifiers such as KNN, SVM, NN and Logistic Regression were implemented and corresponding accuracy rates were obtained. Support Vector Machine has the highest accuracy with 99.3%.The proposed method was applied to medical dataset which helped doctors to make more correct decision [17]

Various segmentation algorithms were discussed which includes Naïve Bayes, Hidden Markov Model etc. Proper explanation is given about how and why various segmentation algorithms are used in detection of Lung tumor [18].

#### 1.2 Objective of the Research

The general objective of this research is detection of lung tumor and classification its type. The specific objectives of this research are:

* To detect malignant growth from Computed Tomography (CT) images.
* To enhance low-quality images into higher quality for machine learning.
* To classify the lung tumor cancer using CNN.
* To apply different types of machine learning algorithm.
* To compare the accuracy between different types of algorithm.

#### 1.3 Scope and limitation of the Research

This research addresses a large number of scope and function as this can identify the malignant growth from CT images and also classify the lung tumor. Early recognition of cellular breakdown in the lungs can build the odds of endurance among individuals. Detecting nodule from CT image can me more time consuming than traditional method. As well as classifying the lung tumor we can early detect the cancer stage of the patient. Analyzing CT images this research can be a better guideline.

The main problem of this research work is data. As we used image data, resolution was low. Data finding is another challenge as data availability in Bangladesh is very tough. Besides the number of leveled dataset, we used was not sufficient for calculating accuracy.

#### 1.4 Study Area Profile

This section describes the area profile of this research that provides a brief idea about the selected area.

Lung cancer is a disease in which malignant (cancer) cells form in the tissues of the lung. Cancer is the abnormal growth of body cells which begin to divide without stopping and may form growths called tumors. Cancerous tumors are malignant which can spread intosurrounding tissues by the procedure of metastasis. Among all distinctive sort of malignant growth, lung cancer is the leading cause of cancer death in both men and women. Lung cancer has two main categories, such as small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC).

Early stage detection and treatment of cancer can reduce the mortality rate. Early detection of lung cancer is difficult as it shows symptoms in final stage. Smoking causes most lung cancers, but nonsmokers can also develop lung cancer. Lung cancer screening may help detect the cancer at early stage. Tests are used to screen for different types of cancer when a person does not have symptoms. Lung disease might be seen on chest radiographs and computed tomography (CT) scans. Among all screening methods, CT scans is the most reliable and effective method as it shows detail picture of the lesions and its growth. The determination is affirmed by biopsy which is typically performed by bronchoscopy or CT-guidance [19]

Lung cancer is basically classified into four stages: Stage I, Stage II, Stage III and Stage IV depending on severity. Stage I cancer is confined to the lung, Stage II and III cancer is confined within the chest and Stage IV lung cancer has spread from the chest to other parts of the body [19] .Visual interpretation may lead to manual errors and wrong interpretation of the stages of cancer. Therefore, computer aided automated systems are required for proper diagnosis.

#### 1.5 Structure of the Thesis

The thesis is organized as follows. First of all, the entire thesis is equipped into five chapter.

This thesis starts with the introduction chapter which is the background of the research. The study area, the objective of the research, and the scope and limitations of the study area are also included in the first chapter.

Second chapter is the theoretical framework and methodology where previous related works and detailed information of the research have been explained.

Third chapter is the analysis of image processing for detecting nodule and classification of lung tumor.

Then in the fourth chapter monitoring accuracy, loss through the dataset for using different types of algorithm.

The final chapter is recommendation and conclusion.

## Chapter 2

### Theoretical Framework and Methodology

The detail description of the methodology of the research along with the literature reviews has been stated in this chapter.

#### 2.1 Basic Terminologies

The basic terms that have been used throughout this research are lung tumor and its classification. In this research, Computed Tomography (CT) images and labeled lung cancer datasets are used. Matlab is a kind of software that helps in analyzing image processing and python is a kind of software which classify lung cancer and show its accuracy.

#### 2.1.1 Lung Cancer

Cancer is generally a malignant tumor. Lung cancer is a disease caused by unchecked growth and the spread of cells from the lung. Affected people have an abnormal cell in his lung which cluster together and eventually this form a tumor. In a cancer cell, there is no cell division control which destroys healthy tissue around that area. When the lung is affected by cancer the cancer cell grows so fast that it prevents that part of the lung from breathing properly. Sometimes cancerous tissue can spread into different parts of the body through the blood or lymph. The American Cancer Society estimates that in 2021 there will be 235,760 new cases of lung cancer in which 119,100 men and 116,660 women and 131,880 deaths from lung cancer in the United States. Lung cancer is the leading cause of cancer death which is around 25% of all cancer deaths. This high mortality rate occurs as the early stage cancer is asymptotic in nature and detected when the cancer is already locally advanced or has disseminated.

Lung cancer can be classified into small cell lung cancer and non-small cell lung cancer. Small cell lung cancer can be divided into small cell carcinoma and combined small cell lung cancer. Non-small cell lung cancer can be divided into Adenocarcinoma, Squamous cell carcinoma and large cell carcinoma.

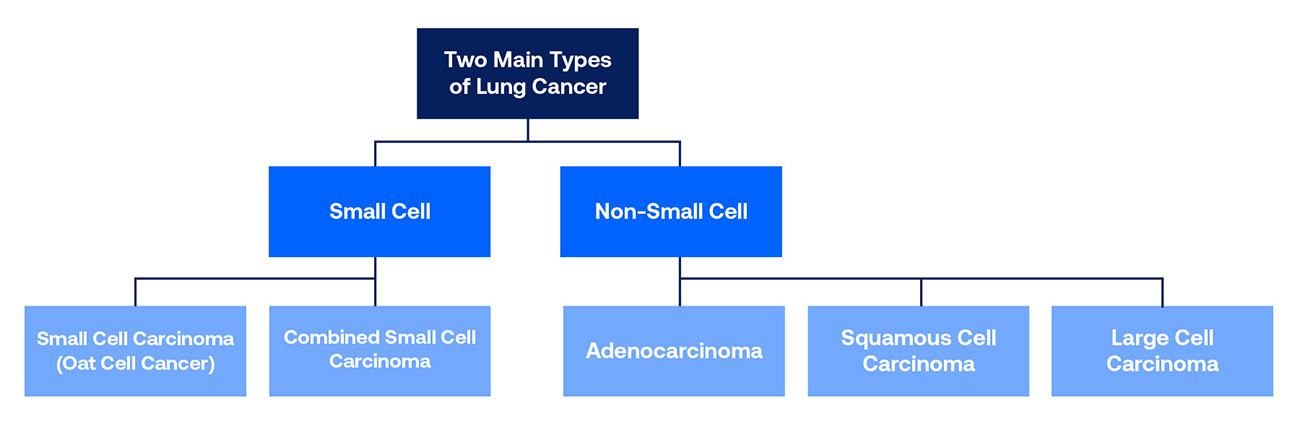


Figure 2.1: Lung Cancer Type

#### 2.1.2 Lung Tumor Classification

Lung tumor can be classified into two terms such as Benign and Malignant.

A benign lung tumor a strange development of tissue that fills no need and is discovered not to be destructive. Generous lung tumors may develop from a wide range of designs in the lung. Deciding if a knob is a favorable tumor or a beginning phase of malignant growth is vital.

Malignant tumors are carcinogenic. They create when cells develop wildly. On the off chance that the cells proceed to develop and spread, the sickness can become dangerous. Threatening tumors can develop rapidly and spread to different pieces of the body in an interaction called metastasis.

#### 2.2 Literature Review

Smoking is said to be the main cause of having lung cancer. But non-smokers can also develop lung cancer. Other than that, American Cancer Society has mentioned many more risk factors that can develop lung cancer such as tobacco/marijuana/e-cigarettes smoking, second-hand smoking, exposure to radon and uranium (radioactive elements), exposure to asbestos, personal and family history of lung cancer, etc .Lung cancer has basically four stages (Stage 1, Stage 2, Stage (3A, 3B), Stage 4) depending on the extremity of the cancer. Interpretation of medical images manually can be erroneous and detect wrong stage of cancer. Inaccurate detection may cause faulty treatment which can be hazardous for patient’s life besides increases the mortality rate. Thus, the automated system can be efficient to develop this consequence. There have been conducted several types of researches for tumor detection and prediction of cancer stages in recent years. Different approaches of image processing techniques and deep learning models have been implemented to develop this sector. In [20], the authors use SVM (support Vector Machine) and ANN (Artificial Neural Network) to classify 250 lung images into two groups (Normal and Abnormal). They use Global Threshold Technique for image segmentation and Ant Colony Optimizer (ACO) for feature selection. The system has better accuracy of 98.40% using ACO\_ANN compared to ACO\_SVM. In [21], the authors use both supervised machine learning strategies using 3D convolutional neural network and unsupervised machine learning strategies using SVM to classify Benign and Malignancy of the cases with 91% accuracy. In [22], the authors proposed fuzzy particle swarm optimization (FPSO) method with Convolutional Neural Network on lung cancer images to achieve 99.2% accuracy to classify Benign and Malignant case. In [23], the authors demonstrated three machine learning predictive models to diagnose cancer using description of nuclei sampled from breast masses. They used supervised learning such as General Linear Model regression (GLMs), Support Vector Machines (SVMs) with a radial basis function kernel, and single-layer Artificial Neural Networks where they have achieved 96% accuracy and area under the curve of 97% using SVM. In [24], authors proposed a method for early detection of lung cancer using pre-processing of images using image processing tools, feature extraction using Gabor Filter and K-NN classification using Genetic Algorithm where K (50-100) number of samples were picked for single iteration. They found 90% accuracy using this algorithm. In [25], the authors have implemented five classification models such as Decision Tree(DT), K-nearest Neighbor (KNN), Support Vector Machine(SVM), Ensemble Tree(ET) and Back Propagation Neural Network(BPNN) in two stages :- Detection of the tumor and Staging. Average accuracy of 92.8% for detection and 90.6% for staging has been achieved using BPNN model. They have only focused on T-stage cancer classification in their research. In [26], the authors proposed a new algorithm for feature extraction using image processing and implemented seven different supervised machine learning algorithms to classify images into Benign and Malignant classes. They have found highest accuracy of 88.55% using MLP (Multi-Layer Perceptron) classifier algorithm in compare to the other classification algorithm. In [27], the authors implemented median filter for de-noising and morphological operations for segmentation of 216 of lung images. ANN was used as the classifier and they achieved 92% of testing accuracy. In [28], authors have proposed a classifier which is based on Convolutional Deep Neural Network. They have implemented the model into previously classified Magnetic Resonance Imaging by medical specialists to detect whether the patient has cancer or not. Hence they have decided to take the result as accurate if the certainty of cancer is 75%. In [29], several image processing methods have been done by the authors to find the ROI (Region of Interest) from the lung CT images and some important parameters have been collected from the extracted tumor. After that, they have implemented SVM (support vector machine) to classify those processed images into two classes (Benign and Malignant). In [30], authors have contributed to detect Parkinson’s disease at early stage on SPECT (Single Photo Emission **Figure 1 shows the system flow chart of our proposed model** Tomography) images using image processing and ANN. They have pre-processed the SPECT images using several pre-processing technique, segmented the ROI (Region of interest) classify the extracted features and achieved 94% of accuracy. In [31], authors have attempted to implement different machine learning algorithms for lung cancer classification. They has been concluded the research with the result that, SVM gives better accuracy of 98.10% for classifying lung cancer into different stages compared to other classifier discussed on the paper. Different authors have contributed a lot to process cancer images and classify them into several classes but classification of the tumor as Normal, Benign and Malignant has not been implemented.

#### 2.3 Methodology of the Research

The steps followed to accomplish the targets and do the whole research have been described in this section.

The following research have been divided into two methods.

First one is detection of lung nodule, image pre- processing and classification from CT images using MATLAB.

Second one is lung tumor classification using different types of machine learning algorithms for better accuracy.

#### 2.4 Detection of nodule using MATLAB

To extract the features and to detection the nodule we have chosen CT scan image for carrying out this research. As the research is based on secondary data, Data has been collected from cancer imaging archive (TCIA).

#### 2.4.1 Proposed methodology

We propose a system which can detect lung tumor region using the CT scan image processing and can classify the lung disease as Benign or Malignant using machine learning. Figure 2 shows the system flow diagram of our proposed system model. The system works in five principle steps: Image pre-processing, Image segmentation, Training features extraction, Test feature extraction and Classification.

Input Lung Tissue Test Image

Image Pre-processing using Median filter and Image Enhancement

Image Segmentation of Lung tumor using Thresholding and Edge Detection

Figure

Geometrical Test Feature Extraction (Area, Perimeter, Eccentricity, Compactness, Circularity)

Classification using Support Vector Machine (SVM)

Training Feature Extraction of the trained Benign and Malignant Tumor

Classified Lung cancer (Benign or Malignant)

fig:1

Figure 2.2: Proposed System Model

First, the system takes the input image for testing. Then, the primary step start with pre-processing of the images which applies a few procedures of picture resizing, smoothing and sharpening to get the best degree of value and clearness. The subsequent step applies segmentation of the lung tumor using thresholding and edge detection methods. Then the features of the trained benign and malignant tumors are extracted. The next step extracts five features of the segmented tumor region of the test image. The features are Area, Perimeter, Eccentricity, Compactness and Circularity. The final stage classifies the test features as Benign and Malignant using support vector machine (SVM) and gives the classified output as a marker.

#### 2.4.2 Methodologies

To develop the system, first we collect data from the accessible Cancer Imaging Archive database. The database contains an assortment of Computed Tomography (CT) images of patients with and without lung cancer. We collect CT images of patients with lung disease (Benign and Malignant) from the database. The CT images are in DICOM format. The images are used as the training and test images to develop the system. The methodologies are explained as below:

#### A. Image Pre-Processing

The image pre-processing aims to remove unwanted noise and enhance valuable features of the image. It includes three steps as image resizing, image smoothing, and image enhancement.

* Image Resizing: First, all the CT images are resized to 256 x 256 dimensions. Then, we crop the images to get a specific position and dimension.
* Image Smoothing: We use median filtering for image smoothing. It is a non-linear operation that is powerful than convolutional operation while de-noising and keeping the edges unchanged. The filter uses a zero-padding technique on the edges so that the median value of the edges remains unaffected [32].
* Image Enhancement: Finally, image enhancement operation mainly sharpens image features such as boundaries, edges and reduces the artifacts caused by contrast variations. For our proposed system, we change the contrast level of the input image. We use a tool specially made for image adjustment in MATLAB to adjust the intensity level. This tool helps to portray the intensity values of a grayscale image into a new valued image. By default, it saturates few data at high and low intensities which helps to enhance the output image [33].

After the image pre-processing operation, we can check the contrast by finding and displaying the histogram of the image data. The histogram plot uses 256 equally spaced bin, each represents a range of data values. We have to smooth the image to produce a less pixelated image if the contrast level is not satisfactory.

#### B. Image Segmentation

Image segmentation is the process of partitioning an image into multiple segments. The goal of segmentation is to represent the image into something that is easier to analyze. Generally, it is used to locate any object or boundaries in the image. Several general-purpose algorithms and techniques have been developed for image segmentation. The effectiveness of the techniques depends on the specific problems or application requirements.

The goal of our image segmentation methodology is to locate and segment the lung cancer nodules or the region of interest (ROI) of the lung cancer in the lung tissues. We apply Thresholding and Edge detection followed by contrast adjustment and clear border operation for segmentation of the lung tumors.

* Thresholding: Thresholding method is based on a clip-level (or a threshold value) to turn a gray-scale image into a binary image. We use Thresholding method to compute the threshold or a level. The value of the threshold level is applied to change the grayscale images to binary images [34]. Here the level indicates the intensity value of the image. Several popular Thresholding methods are used in industry including the maximum entropy method, balanced histogram thresholding, Otsu's method (maximum variance) [35], and k-means clustering. Otsu’s method is used to compute global threshold from grayscale image. This method selects the threshold which is used to minimize the intra-class variance of the black pixels and white pixels [36].It ignores any nonzero.
* Edge Detection: The edge detection process finds the edges of the background binary image. There are different edge detection methods as sobel, prewitt, roberts, canny edge etc. We detect the edges of the binary images using ‘Roberts’ method as it performs best for binary images. The detected edge is then subtracted from the binary image. We apply contrast adjustment to adjust the intensity values of the edge subtracted binary image. Finally, clearborder operation is performed on the enhanced binary image to suppress the light structures connected to the border and to segment the ROI of the lung tumor from the lung mask.

D. Classification: This is the final stage where the detected tumor is classified as malignant or benign. We apply Support vector machine (SVM) as classifier. Figure 2 shows the classification process.

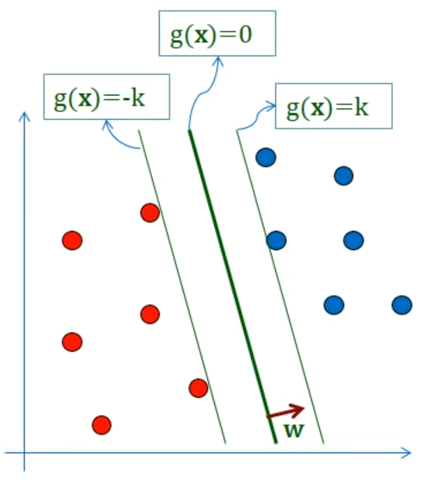


Figure 2.3: SVM Classifier

It is supervised machine learning method that analyze input data and classify them according to pattern. SVM defines the group or classifier function using training datasets and classify them into two classes or groups. Then it assigns the test datasets into one of these two groups. The predicted group function is defined as

……………..2.1

where xi are training inputs, wT is m dimensional vector, and b is bias term. Here, i=1…. M. The value of g(x) depends on ||w||.

……………….2.2

……………2.3

#### 2.5 Classification of lung tumor using deep learning and machine learning algorithms:

Further to get better accuracy we have chosen labeled data from Kaggle [37] for better accuracy. For this we used here both processed and unprocessed dataset for comparing result. Different types of algorithms we have used here such as

* CNN
* VGG16
* KNN
* Decision Tree
* Random Forest classifier
* Support vector machine
* Naive Bayes algorithm

#### ****2.5.1 Image Processing:****

**The main objective of processing the images is to get better accuracy in the system. We have focused to make the lung images noise-free, increased the intensity and enhanced the images to a certain level for better computer vision, segmented the lungs better training of the machine. The image processing part includes some steps; - contrast enhancement, resizing, filtering, segmentation.**

* Image Resizing:**For the resizing purpose at first, we converted the RGB image into a grayscale** **image. Then we resized the images from 207 × 130 to 256 × 256. Then cropping of the images has been done necessarily.**
* Median Filtering: **CT scan images usually contain noises like “salt and pepper” noise. For reducing these noises while preserving the edges, we found the Median filter to be most effective than other filters. By default, the filter took 3 × 3 neighborhoods to do the process.**
* Image Segmentation: **as mentioned before, our main focus is to enhance the lung images and segmented them from the background to improve the accuracy of the model. For segmentation of the lungs from the background, we determined a certain level of the threshold value and convert the images into binary image based on the threshold level.**

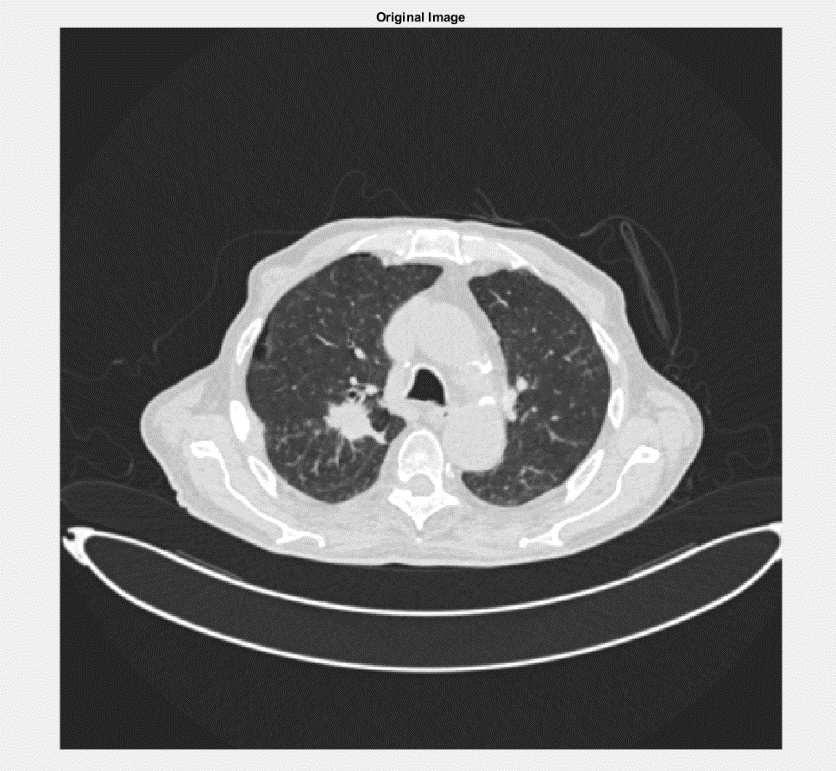
**Some example is given below:**

Figure 2.4: Example of Image Segmentation

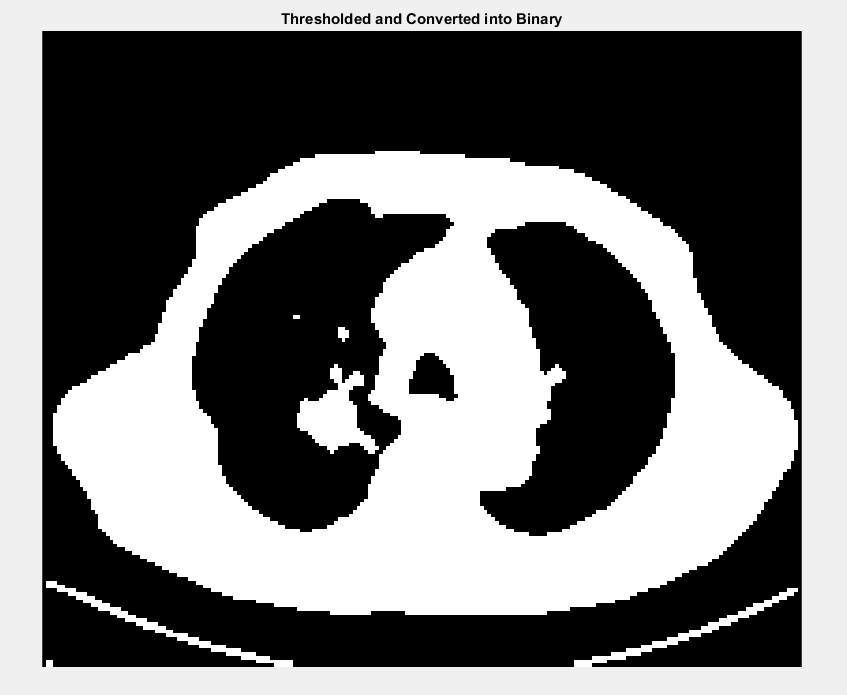
****

Figure 2.5: Example of Image Segmentation

## 

## Dataset Visualization

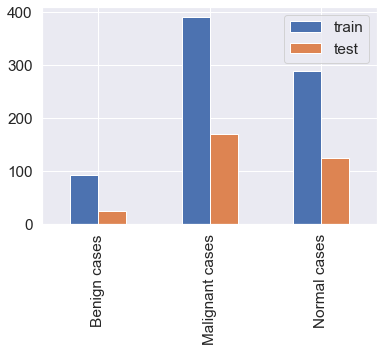


Figure 2.6: Dataset Visualization

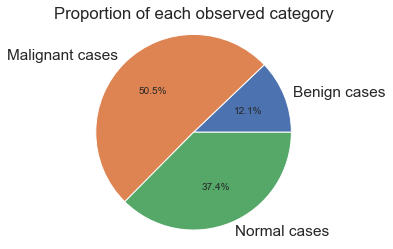
****

Figure 2.7: PI Chart of Data Visualization

From the data visualization we can see that, our dataset has low quantity benign case images. For training we used 768 images for training and 329 images for testing.

2.6 Machine learning: Machine learning takes AI software a step further as it enables intelligent learning to occur within the component based on previous work it did or extrapolations made from data. The software performs sophisticated decision making processes as it goes along and learns from previous activities. A brief description of the research papers based on Lung Cancer detection using different Machine learning algorithms are explained below:

#### 2.6.1 CNN

A CNN is type of a DNN consists of multiple hidden layers such as convolutional layer, RELU layer. Pooling layer and fully connected a normalized layer. CNN shares loads in the convolutional layer lessening the memory impression and expands the exhibition of the organization. The significant highlights of CNN lie with the 3D volumes of neurons, nearby availability and shared loads. An element map is created by convolution layer through convolution of various sub districts of the info picture with a learned bit. At that point, anon-straight actuation work is applied through ReLu layer to improve the union properties at the point when the mistake is low. In pooling layer, a locale of the picture/include map is picked and the pixel with greatest esteem among them or normal qualities is picked as the delegate pixel so a 2x2 or 3x3 matrix will be diminished to a solitary scalar worth. This outcomes an enormous decrease in the test size. At times, conventional Fully-Connected (FC) layer will be utilized related to the convolutional layers towards the yield stage. In CNN design, as a rule convolution layer and pool layer are utilized in some blend. The pooling layer typically completes two sorts of tasks viz. max pooling also, implies pooling. In mean pooling, the normal neighborhood is determined inside the element focuses and in max pooling it is determined inside a limit of highlight focuses. Mean pooling decreases the mistake brought about by the neighborhood size limit and holds foundation data. Max pooling decreases the convolution layer boundary assessed blunder brought about by the mean deviation and thus holds more surface data. Figure 3 shows the engineering of CNN.

Figure 2.: Convolutional Neural Network Model

Dense layer

Convolution +Max Pooling layer

Convolution +Max Pooling layer

Convolution +Max Pooling layer

Convolution +Max Pooling layer

2.6.2 Vgg16: We begin from the VGG-16 [37] network, that originally designed for large-scale natural image classification. VGG-16 has thirteen convolutional layers and three FC (fully connected) layers. The convolutional layers area unit denoted as conv- during this paper, the target dataset is relatively little and therefore the pre-trained VGG-16 is powerful in several segmentation tasks (The pre-trained model is obtained from the coaching of the ImageNet large-scale dataset). Therefore, the transfer learning is employed within the coaching in our paper. as a result of we've changed the VGG-16 network, we tend to solely learn conv-, convolution kernel is 3 × 3, and use maxpooling. During this network, we tend to fine-tuned the VGG-16 network. consistent with [38], expanded convolution will increase the sensation field of the convolution kernel whereas keeping the amount of parameters constant, and additionally ensures that the dimensions of the output feature map remains unchanged. we tend to modified the convolution of conv- to expanded convolution, convolution kernel is 3 × 3, expanded rate is a pair of, and therefore the pooling layer when cov-43 and cov-53 is canceled. we tend to born-again the last 2 FC layers into convolution filters, renamed cov-6 and cov-7, convolution kernel is 7 × 7, expanded rate is four, and additional them to feature sets which will be collective into our multi-scale hyper column descriptors. Following, we tend to build predictor supported multiscale options extracted from multiple layers. thanks to a robust correlation between adjacent layers, actually, there's no have to be compelled to take into account all the layers. we tend to use skip-connections to extract hyper column options from with on-demand interpolation. Next, we tend to learned a few nonlinear predictor for classifying pixels, that is enforced as a multilayer perceptron (MLP) outlined on a hyper column options. We use MLP, which may be enforced as a series of “Fully Connected” layers, followed by the ReLU activation perform. The structure of our network is shown in Figure nine.

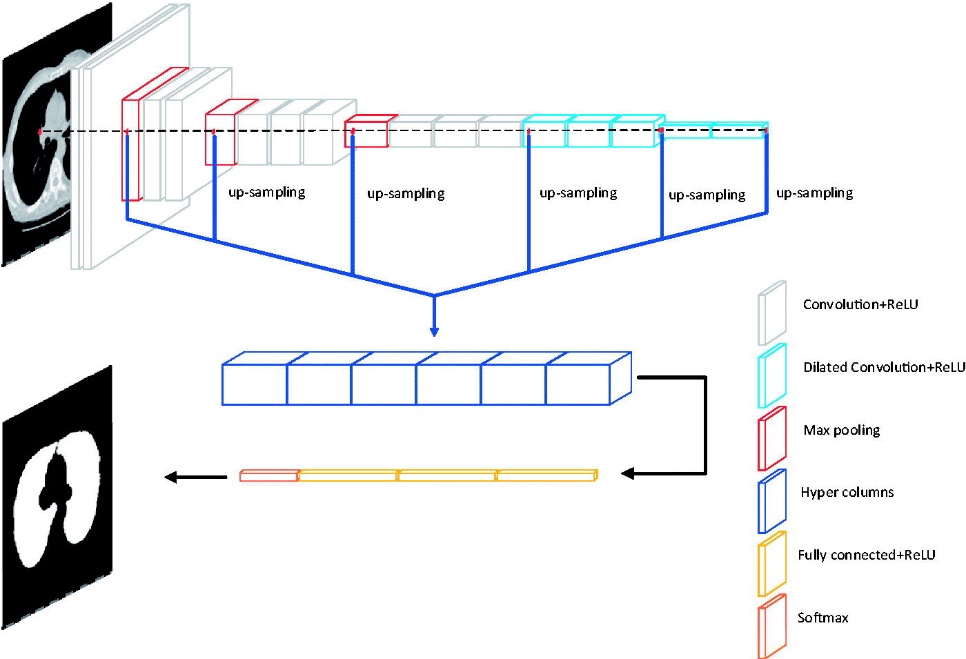


Figure 2.9: Network Structure

Here we first import the VGG16 model from tensorflow keras. The image module is imported to preprocess the image object and the preprocess\_input module is imported to scale pixel values appropriately for the VGG16 model. The numpy module is imported for array-processing. Then the VGG16 model is loaded with the pretrained weights for the ImageNet dataset. VGG16 model is a series of convolutional layers followed by one or a few dense (or fully connected) layers. Include top lets you select if you want the final dense layers or not. False indicates that the final dense layers are excluded when loading the model. From the input layer to the last max pooling layer) is regarded as **feature extraction part**of the model, while the rest of the network is regarded as **classification part**of the model. After defining the model, we need to load the input image with the size expected by the model, in this case, 224×224. Next, the image PIL object needs to be converted to a NumPy array of pixel data and expanded from a 3D array to a 4D array with the dimensions of [samples, rows, cols, channels]*,* where we only have one sample. The pixel values then need to be scaled appropriately for the VGG model. We are now ready to get the features.

#### 2.6.3 K-Nearest Neighbors Classifier

One of the most straightforward grouping methods is the k-closest neighbor (k-NN) classifier. Arrangement of the info include vector X is finished by deciding the k nearest preparing vectors as per a reasonable distance metric. Vector X is then doled out to that class to which the larger part of that k-closest neighbors have a place. The k-NN calculation depends on a distance work and a casting a ballot work in k-closest neighbors; the measurement utilized is the Euclidean distance measure . The k-NN classifier is a customary nonparametric directed classifier that is said to yield great execution for ideal upsides of k. Like most learning calculations, k-NN calculation comprises of a preparation stage and a testing stage. Information focuses are given in a n-dimensional space in the preparation stage. The marks related with the information focuses assign their class in the preparation stage. In the testing stage, unlabeled information are given and the calculation produces the rundown of the k-closest (effectively characterized) information focuses to the unlabeled point. This classifier returns the class of most of that rundown.

#### 2.6.4 SVM (support vector machine)

SVM is an administered learning strategy that break down information which is utilized for grouping investigation. For non-directly distinguishable datasets, SVM is more appropriate since it lessens the misclassification rate. In SVM given an information, the goal is to track down the base separated point from the classes and attempting to discover the expanded distance. Fig. 2.10 shows the design of SVM. Here, green and pink pictures address two unique classes which is isolated by a hyperplane. Likewise the edge and backing vectors are appropriately named underneath [38].



Figure 2.10: Support Vector Machine

#### 2.6.5 Decision Tree

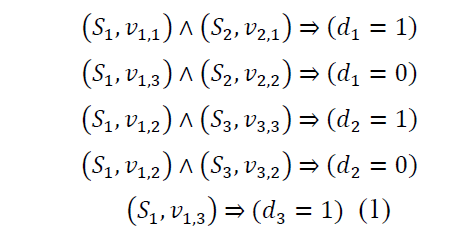
Decision tree utilizes regulated learning method to construct a model which is as a tree information structure (set of hubs organized in various leveled style).Initially, entropy of parent is determined. At that point Information acquire is determined by deducting weighted amount of entropy of kids from entropy of parent. The one with most noteworthy Information acquire is considered as the root hub and the interaction goes on until the arrangement is done. Given another test information , the tree is utilized to foresee the outcome. In decision tree, every hub indicate a specific side effect from the set S ={s1,s2,s3… sj} where S determine restrictive credits, vi, kdenotes the upsides of each branch for example the h-th range for I-th manifestation and leaves which present choices D={d1,d2,… .dk} and their parallel qualities, wdk={0,1} . By recording every way from the root to the leaves,a bunch of affiliation rules was made by changing over the choice tree . [36]

Fig.2.11 describe Decision tree as a bunch of affiliation rule.



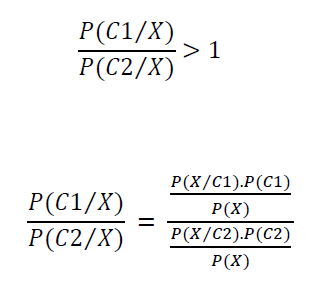
Figure 2.11: GENERAL VIEW OF DECISION TREE

The set of association rules for the above given tree are:

……………2.4

#### 2.6.6 Naive Bayes

Naïve Bayes is mostly used in the area of Data Mining and Machine Learning. Taking advantage of statistical methods SVM classification process is done. We used following equation to calculate the probabilities.

………….2.5

Initially, in order to decide which class the instance belongs to, probabilistic value is calculated. The final class label is the class with highest probability value. In figure 2.12 shown below , there is a new incoming X and each of C1, C2, C3 labels represents the classes. According to probability values given in the figure, class C1 has the highest probability value and therefore the incoming X belongs to class C1 [38].



Figure 2.12: Example X data and Probability-Class Relations

#### 2.6.7 Logistic Regression

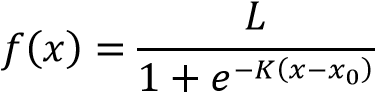
Logistic Regression (LR), a mainstream numerical displaying method utilized in the examination of epidemiologic datasets, particularly space of AI. Calculated Regression strategy can be run in these means:

1. Calculate using calculated capacity.

2. Learn the coefficients for a calculated relapse model.

3. Finally, make forecasts utilizing a calculated relapse model.

The strategic capacity is given beneath:

 ……….. (2.6)

E= Euler's number

x0= Middle x-value of sigmoid function

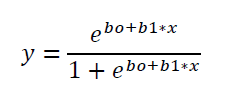
L= the maximum value of curve

K= Abruptness of curve.

Input values (x) to estimate an output value (y);

Logistic regression equation is used.

The logistic regression model is given in equation as:

(2.7)

Logistic regression parameters are estimated by maximizing logarithmic likelihood function using training data. Fig 2.13 shows the example of a logistic regression to distinguish two classes (orange- yellow images).[36]



Figure 2.13: Logistic Regression to distinguish two classes

#### 

#### 

#### 2.7 System Architecture

Our proposed framework followed data acquisitions, data formatting, model training, testing, prediction and comparison result . Figure 2.14 shows the methodology diagram.

Data Acquisition

Data Formatting

Model Training & Testing

Vgg-16 feature extraction with Machine learning algorithms in processed or without processed dataset.

Machine learning algorithms with or without feature extraction

Vgg-16 feature extraction with CNN in processed and without processed dataset

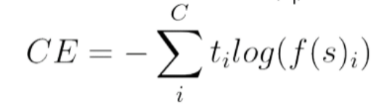
CNN with processed and without processed dataset

Classification

Prediction and comparison Result

Figure 2.14: System Architecture

* **Data Acquisition:** The CT scan images are collected from kaggle dataset. Three classes of benign images, normal images, and malignant images of lungs with 1097 images are considered for our work.
* **Data Formatting:** The obtained dataset was RGB color CT scan pictures with .jpg format. The pictures were resized to keep a uniform perspective proportion of one with (224,224) pixel size for the deep learning and machine learning activity. For data augmentation first we rescale the image to transform every pixel value from range [0,255]. Then by shearing, zooming the images 30% and also by doing horizontal and vertical flip we can extract more data from the images.
* **Model Training & Testing:** A liner stack of layers was used to create the deep learning and machine learning model for the image classification and recognition. Training and testing images were passed through convolutional layers or vgg-16 feature extract with kernel filters, max pooling, and fully connected layers. The softmax function was applied to classify the given object. The model was trained and tested using Google Colaboratory GPU .A neural network with three hidden layers, one input layer, and one fully connected layer was implemented for this task. Images are split in a ratio of 70:30 for training and validation purposes. Images of (224, 224) pixel size were passed to the input layer. Kernel matrix of (3, 3) with (ReLU(x) = max (0, x)) as an activation function was applied in each convolutional layer. Max pooling size of (2, 2) was implemented to reduce the computation parameters in the next convolution layer. A dense value of three with the sigmoid activation function was used to obtain the class probabilities for final output classes. An adaptive moment estimation (Adam) optimizer was used to calculate the learning rates for different parameters. Loss function calculates the discrepancy between the predicted output and the labeled output for the given input; sparse categorical cross-entropy (CE) was used as a loss function for this task, which is calculated as:

…. (2.8)

C is the number of output class, Sp is the accuracy score of the given positive class, and tj is the score inferred by the net for each class C.

For beter accuracy we have run our model CNN with machine learning algorithms,

For CNN with processed or without processed data epoch=20 and batch size=64.For vgg-16 feature extraction with processed or without processed data epoch=20 batch size=32.

The trained model weights were saved into the hd5 file format and used to predict the future by loading the weights to the model architecture.

**Here is some predicted and unpredicted images which was taken from test process using system Model:**

## CNN with Processed Dataset

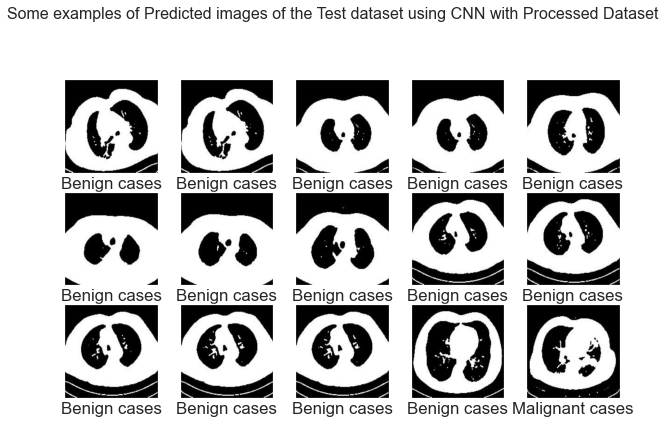
****

Figure 2.16: Example of Processed Image of the Test Dataset Using CNN with Processed Dataset

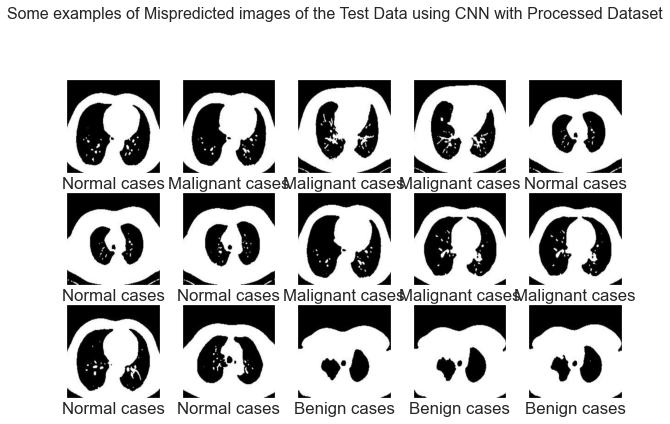
****

Figure 2.17: Example of Mispredicted Images of the Test Data using CNN with Processed Dataset

## CNN without Processed Dataset

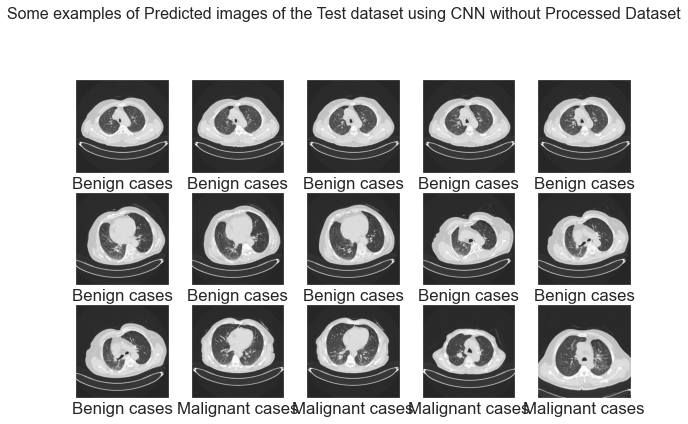


Figure 2.18: Examples of Predicted Image of the Test Dataset using CNN without Processed Dataset

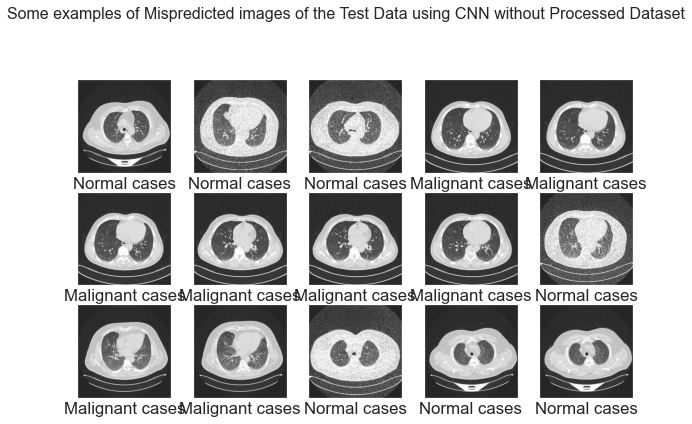
****

Figure 2.19: Example of Mispredicted Images of the Test Data using CNN without Processed Dataset

## Support Vector Machine on processed Dataset

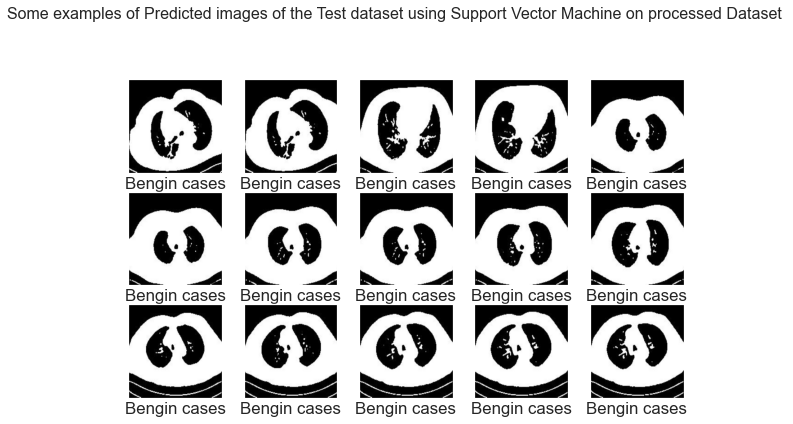


Figure 2.20: Example of predicted Images of the Test Data using SVM ont Processed Dataset

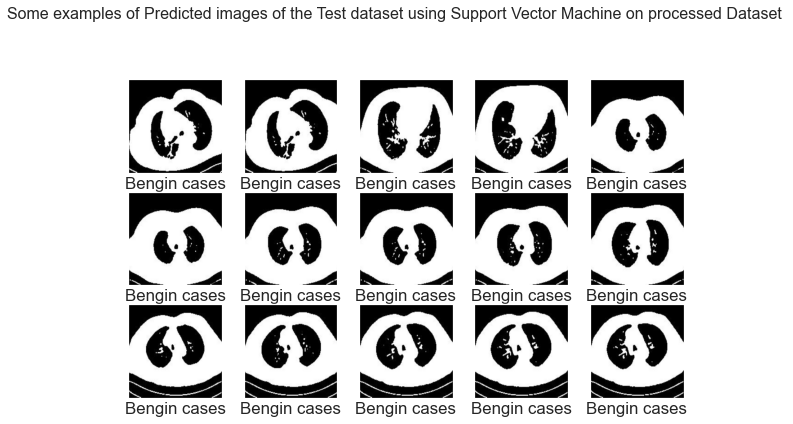
****

Figure 2.21: Example of Mispredicted Images of the Test Data using CNN without Processed Dataset

## Logistic Regression Classifier on processed Dataset

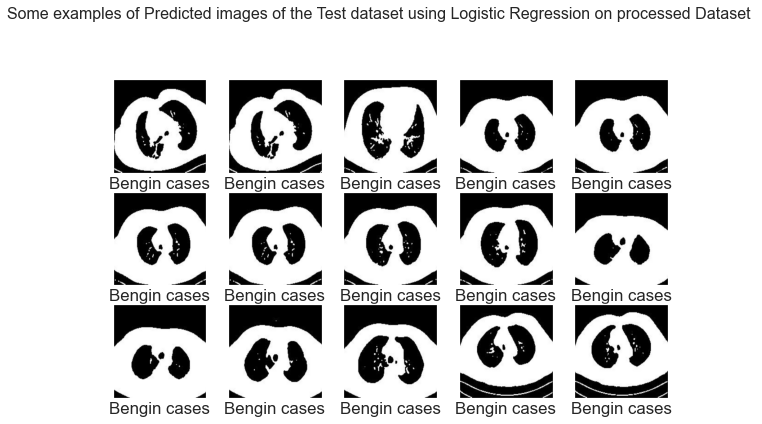
****

Figure 2.22: Example of predicted Images of the Test Data using Logistic Regression Classifier on Processed Dataset

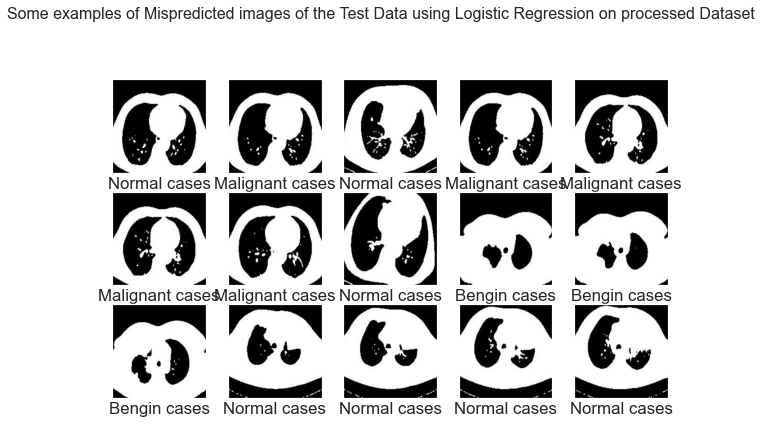
****

Figure 2.23: Example of Mispredicted Images of the Test Data using Logistic Regression Classifier on Processed Dataset

## Random Forest Classifier on processed Dataset

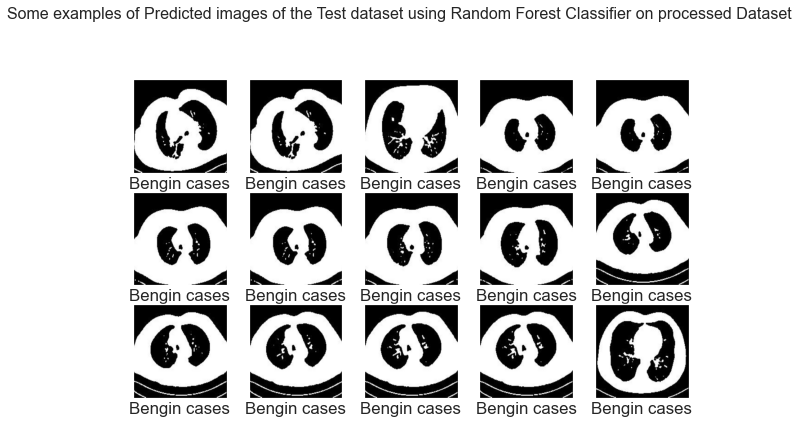


Figure 2.24: Example of predicted Images of the Test Data using Random Forest Classifier on processed Dataset



Figure 2.25: Example of Mispredicted Images of the Test Data using Random Forest Classifier on processed Dataset

## Unprocessed Dataset

## Support Vector Machine on Unprocessed Dataset

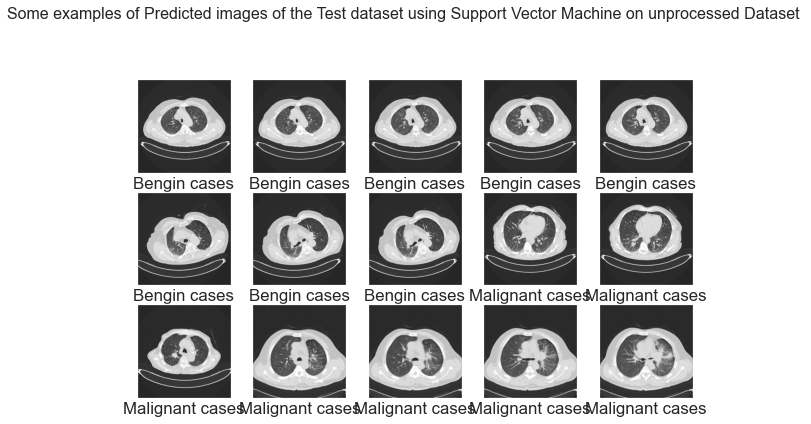


Figure 2.26: Example of predicted Images of the Test Data using Support Vector Machine on Unprocessed Dataset

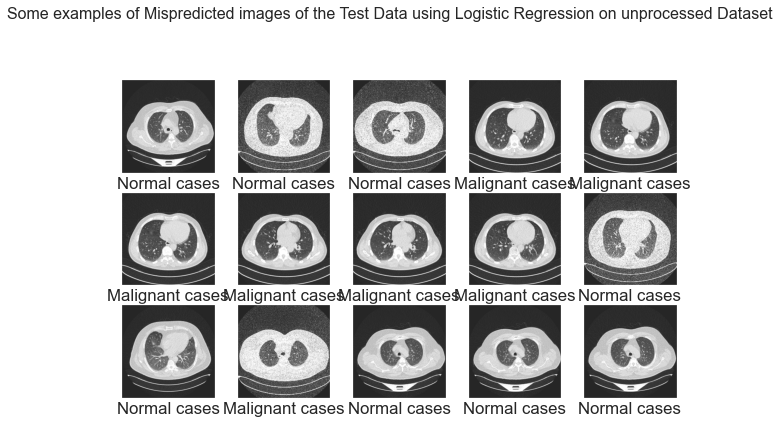
****

Figure 2.27: Example of Mispredicted Images of the Test Data using Support Vector Machine on Unprocessed Dataset

## Logistic Regression on unprocessed Dataset

****

Figure 2.28: Example of predicted Images of the Test Data using Logistic Regression on unprocessed Dataset

Figure 2.29: Example of Mispredicted Images of the Test Data using Logistic Regression on unprocessed Dataset

## Random Forest Classifier on unprocessed Dataset

****

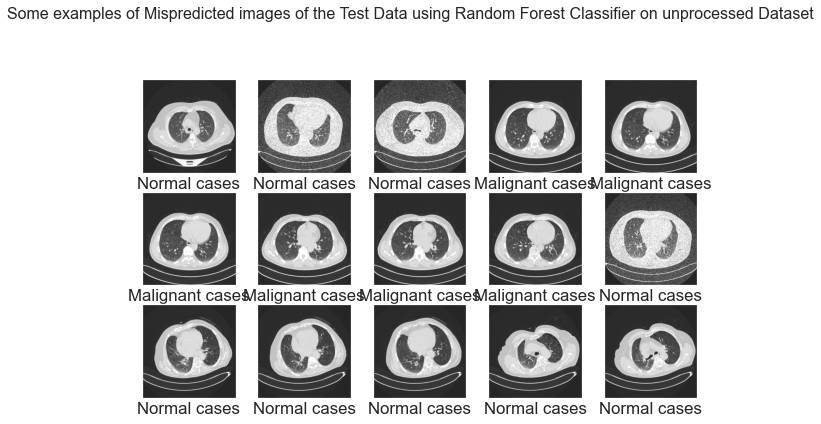
Figure 2.30: Example of predicted Images of the Test Data using Random Forest Classifier on unprocessed Dataset

Figure 2.31: Example of Mispredicted Images of the Test Data using Random Forest Classifier on unprocessed Dataset

# Vgg16 feature extraction with Processed Dataset

## Support Vector Machine on Vgg16 feature extraction with Processed Dataset

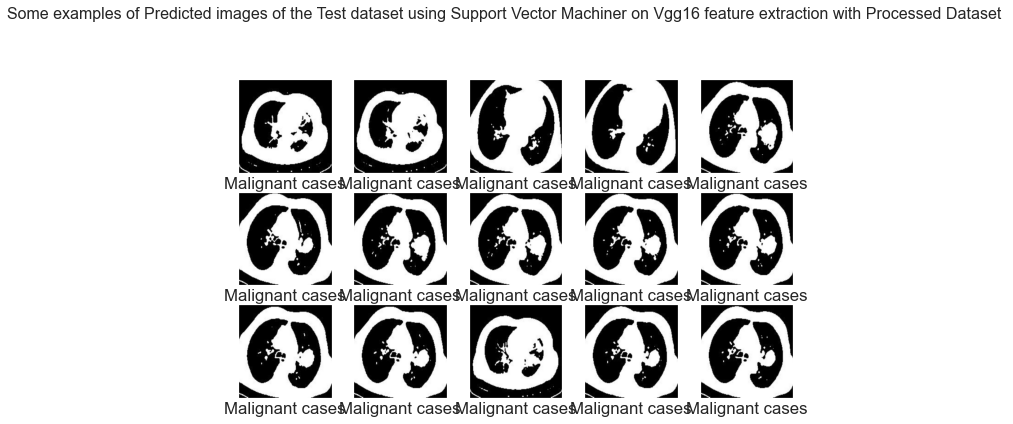
****

Figure 2.32: Example of predicted Images of the Test Data using Support Vector Machine on Vgg16 feature extraction with Processed Dataset

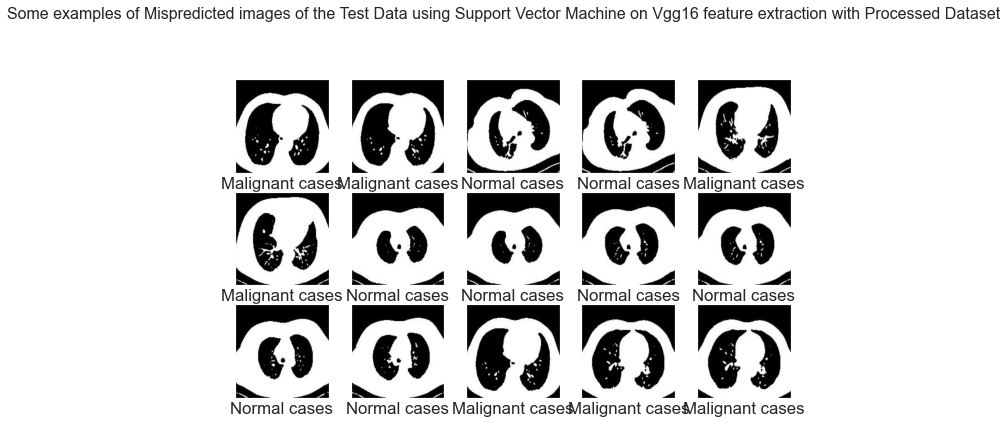
****

Figure 2.33: Example of mispredicted Images of the Test Data using Support Vector Machine on Vgg16 feature extraction with Processed Dataset

## Random Forest Classifier on Vgg16 feature extraction with Processed Dataset

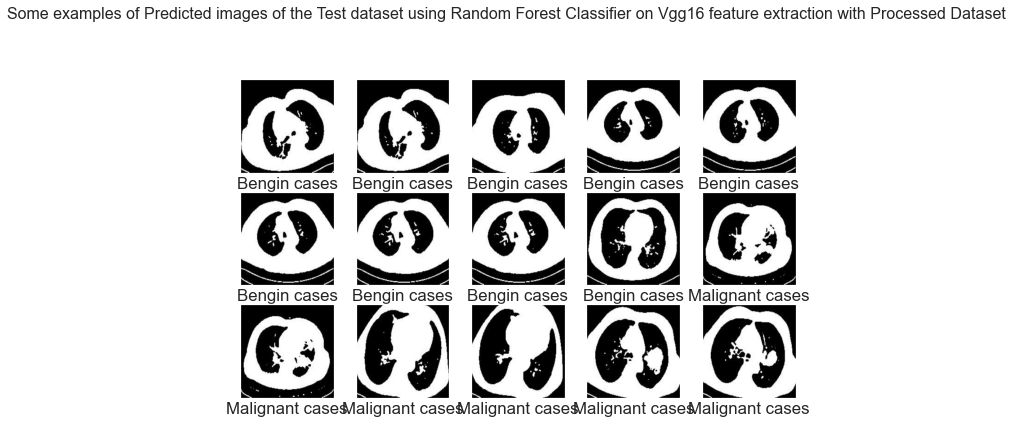
****

Figure 2.34: Example of predicted Images of the Test Data using Random Forest Classifier on Vgg16 feature extraction with Processed Dataset

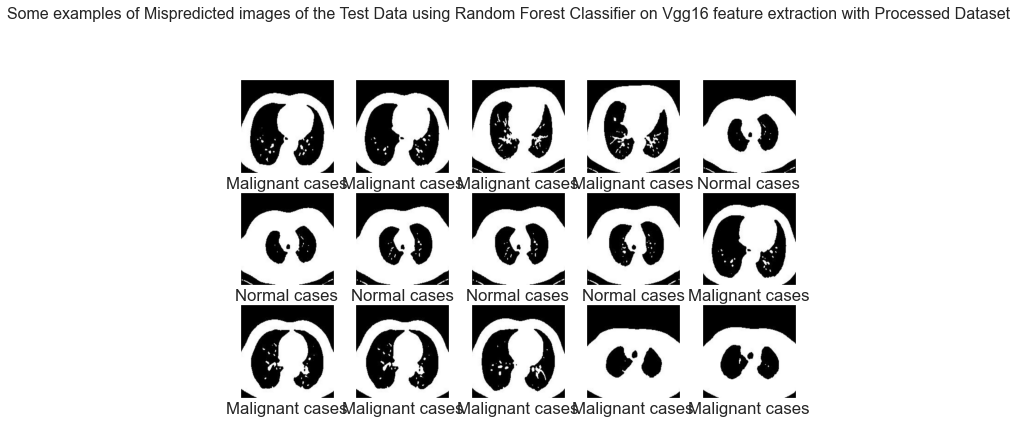
****

Figure 2.35: Example of mispredicted Images of the Test Data using Random Forest Classifier on Vgg16 feature extraction with Processed Dataset

## Logistic Regression on Vgg16 feature extraction with Processed Dataset

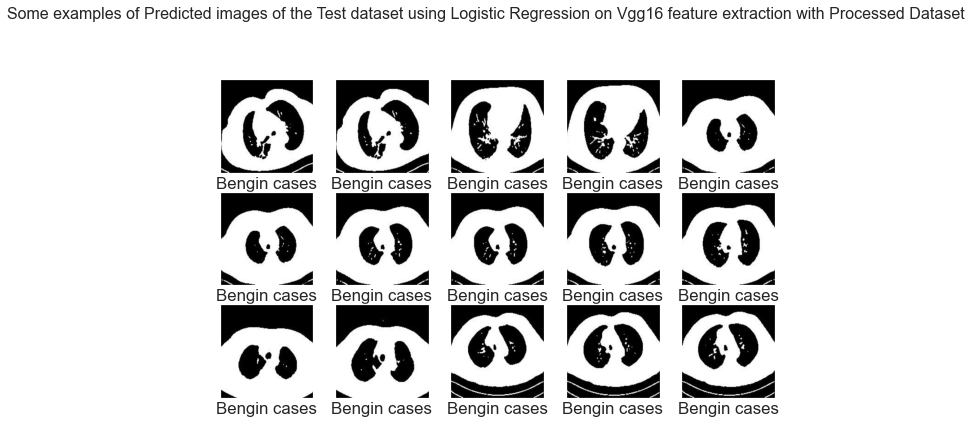
****

Figure 2.36: Example of predicted Images of the Test Data using Logistic Regression on Vgg16 feature extraction with Processed Dataset

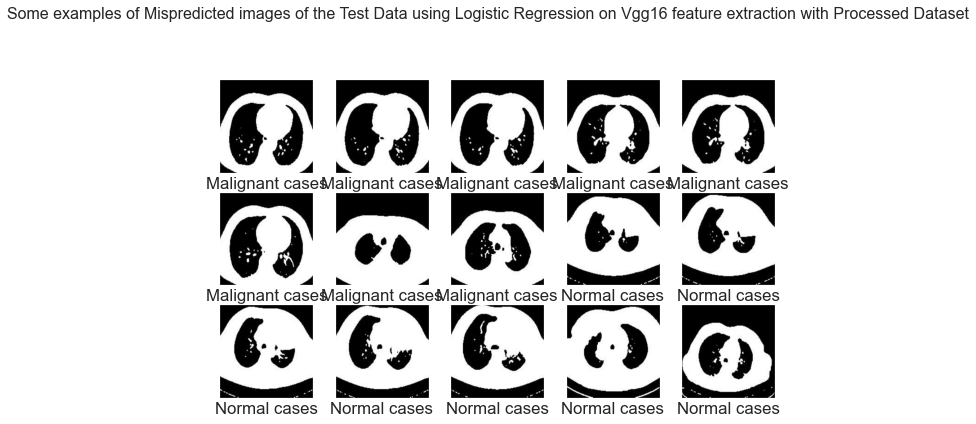
****

Figure 2.37: Example of Mispredicted Images of the Test Data using Logistic Regression on Vgg16 feature extraction with Processed Dataset

## Support Vector Machine on Vgg16 feature extraction without Processed Dataset

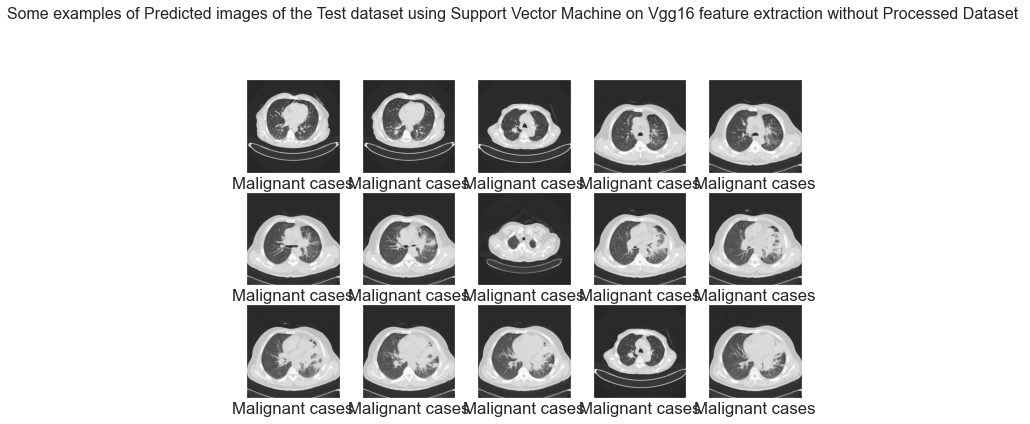
****

Figure 2.38: Example of predicted Images of the Test Data using Support Vector Machine on Vgg16 feature extraction without Processed Dataset

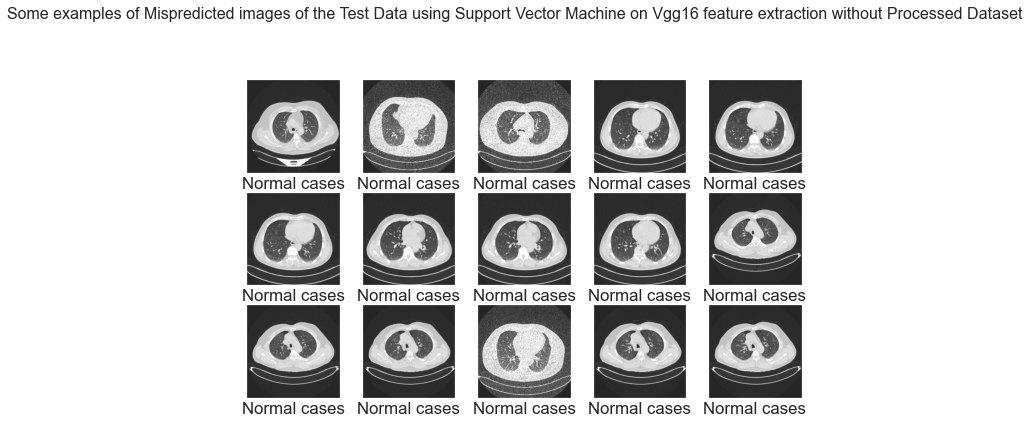
****

Figure 2.39: Example of Mispredicted Images of the Test Data using Support Vector Machine on Vgg16 feature extraction without Processed Dataset

## Logistic Regression on Vgg16 feature extraction without Processed Dataset

****

Figure 2.40: Example of predicted Images of the Test Data using Logistic Regression on Vgg16 feature extraction without Processed Dataset

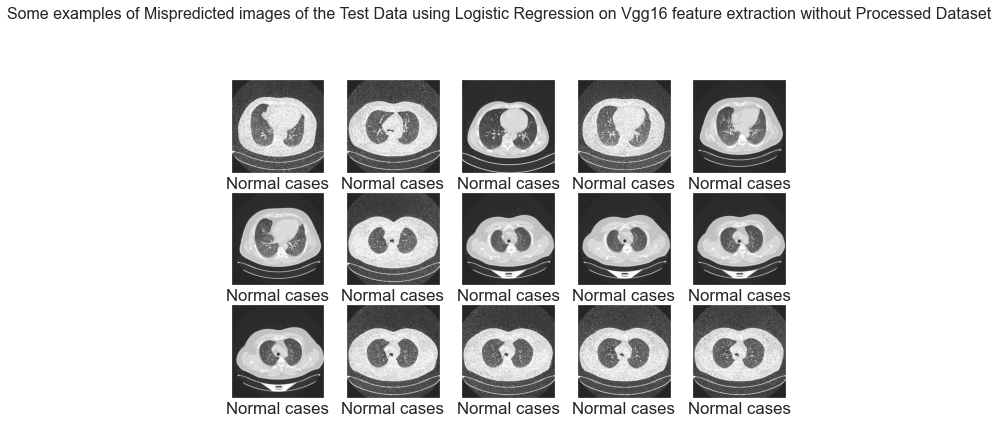
****

Figure 2.41: Example of Mispredicted Images of the Test Data using Logistic Regression on Vgg16 feature extraction without Processed Dataset

## Random Forest Classifier on Vgg16 feature extraction without Processed Dataset

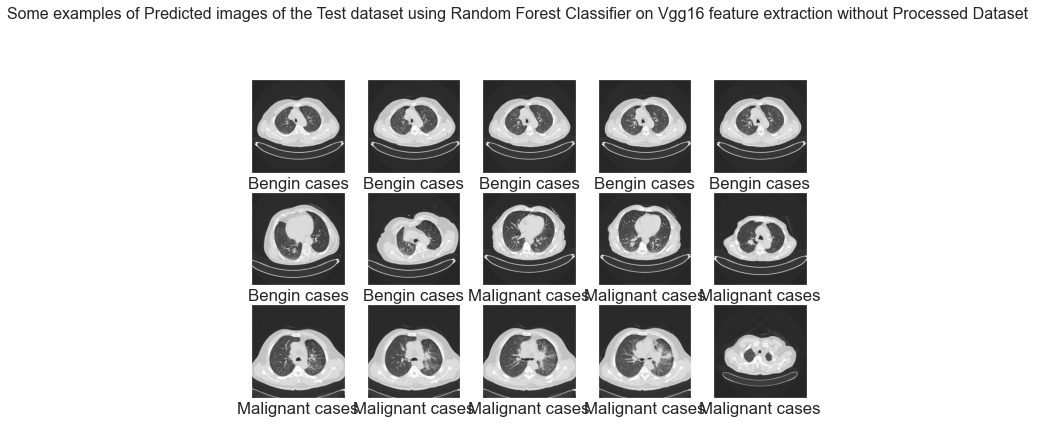
****

Figure 2.42: Example of predicted Images of the Test Data using Random Forest Classifier on Vgg16 feature extraction without Processed Dataset

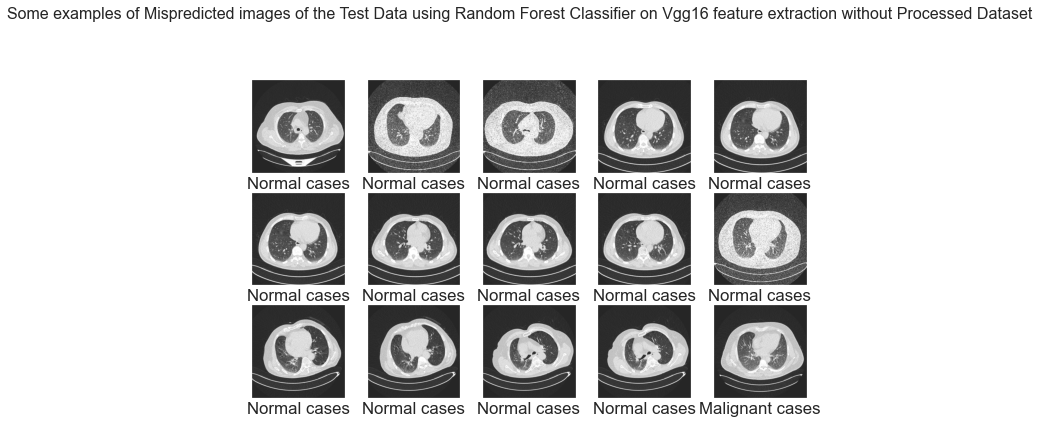
****

Figure 2.43: Example of Mispredicted Images of the Test Data using Random Forest Classifier on Vgg16 feature extraction without Processed Dataset

## Chapter 3

## Monitoring and Analysis of system model

#### 3.1 Feature Extraction from Image processing

Feature Extraction is a significant stage in image processing. Feature extraction is the collection of measured data or features that are informative and calculated from derived values. The size and state of the tumor present in the lungs are assessed by extracting the mathematical features of the tumor ROI. We use regionprops operation to measure all the region properties of the labeled ROI of the tumor. Then we calculate five of the features of the tumor region from the measured properties which are recommended for the classification of Benign and malignant tumor [12, 13]. The extracted features are Area, Perimeter, Eccentricity, Compactness and Circularity. The features are defined as below:

* **Area:** The regionprops operations calculates the area as the actual number of pixels in the region. The resolution of our CT scan images are 300 dpi. As we know, 1 dpi = 1 ppi = 25.8 mm, we can calculate the area in mm as follows:

Thus, by dividing the actual number of pixels in the region by 11.8, we

calculate the area in mm.

* **Perimeter:** This is the distance around the boundary of the region. regionprops computes the perimeter as the number of pixels in the boundary of the region. The perimeter is found by calculating the distance between each adjoining pair of pixels around the border of the region as follows:

**

We divide the measured number of pixels by 11.8 to calculate the perimeter in

mm.

* **Eccentricity:** The ellipse which has an equivalent second-moments as the region specifies the eccentricity. It is the ratio of the distance between the foci of the ellipse and the length of its major axis. This metric value is also referred to as the irregularity index of circularity or roundness. The value is between 0 and 1. An ellipse with 0 eccentricity can be defined as a circle and the ellipse can be defined as line segment if the eccentricity is 1. [20]
* **Compactness:** Compactness is defined as the ratio of the area of an object to the area of a circle with the same perimeter where the circle is used with the most compact shape. In case of lung tumors, benign tumors are more smooth and round shape than malignant tumor. Therefore, a malignant tumor with a number of concavities or spicules could be expected to possess a higher value of compactness than a smooth and round benign mass [13]. In this case, we use the alternate formula of compactness as follows:



In this case, compactness value is zero for a circle, and increases with shape elongation or roughness.

* Circularity: It is a measure of roundness or circularity (area-to-perimeter ratio) which excludes local irregularities. It can be obtained as follows:

………… (3.3)

It equals 1 for a circular object and less than 1 for an object that departs from circularity.

#### 3.2 Monitoring from image processing

The input test images go through the steps of image pre-processing, image segmentation and feature extraction. Figure 3.44 shows two input CT scan images of 2 of patients influenced by lung disease and its respective histograms. Histogram picture is a graphical portrayal of a picture which gives pixel dissemination among its different dark levels.



Figure 3.44: Two of the Lung disease CT images and their Histogram

Further in the preprocessing stage Median filter [3x3] has been used to remove salt and pepper noise. After image resizing and smoothing operation using median filter, we get the cropped image and median filtered image as shown in Figure 3.45.



Figure 3.45: Cropped image and the median filtered image of the CT images

Further improvement is done to enhance the image utilizing contrast adjustment. Figure 3.46 displays the enhanced CT images and their histograms. As we can see that the histograms are much improved after the image pre-processing operations.



Figure 3.46: The Enhanced Images and their Histograms

The segmentation of lung tumor is done using thresholding and edge detection. The image is converted to binary image using the threshold values extracted using the Otsu’s method of thersholding. The segmentation procedures separate the lung tumor from the lung mask. Therefore, the edges of the image need to be detected to subtract it from the binary image. After the segmentation procedures, we get the Edges of the binary image and edge subtracted binary image as shown in Figure 8. Then, after contrast adjustment and performing the clear border operation, we obtain the ROI of the lung tumor for the two CT scan images as shown in Figure 3.47



Figure 3.47: Edges of the Binary images and the Edge subtracted Binary images

By feature extraction, we compute five geometrical features as area, perimeter, eccentricity, compactness and circularity of the training and test CT images. Features are extracted using regionprops operation of MATLAB. Then compactness and circularity is calculated using equation (3.2) and (3.3). Training features are used as the standard features for classification of the diseases into two groups as Benign or malignant tumor.

#### 3.3 Analysis of SVM from Matlab tools

We use SVM to train the training features as benign and malignant tumor. The extracted test features of the detected tumor ROI are then compared with the training features using SVM classifier and assigned to any one of the two groups as Benign or Malignant depending on proximity of values. Figure 3.48 shows classification of the detected lung tumor for the test CT image 1 and 2.



Figure 3.48: ROI of the Lung Tumor and their Classification

We have simulated 168 such CT images using our proposed algorithms of lung cancer detection and classification. The training features of benign and malignant tumor and the classification of the extracted test features of two CT scan image are shown in Table 1.

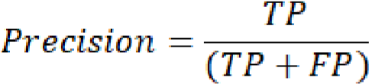
| **ROI Properties** | **Training Features** | | **Test Features and Classification of the Detected ROI** | | | |
| --- | --- | --- | --- | --- | --- | --- |
| ***Benign*** | ***Malignant*** | ***Image 1*** | ***Classification*** | ***Image 2*** | ***Classification*** |
| Area (mm) | >6 | >6 | 26.27 | Benign | 35.42 | Malignant |
| Perimeter (mm) | >6 | >6 | 5.96 | 8.41 |
| Eccen-tricity | 0.11 | 0.8 | 0.62 | 0.84 |
| Compact-ness | 0.3 | 1 | 0.677 | 0.99 |
| Circularity | 0.89 | 0.11 | 0.772 | 0.524 |

Table 1: classification based on training features and the extracted test features

#### 3.4 Classification Report:

A Classification report is used to measure the quality of predictions from a classification algorithm. How many predictions are True and how many are False. More specifically, True Positives, False Positives, True negatives and False Negatives are used to predict the metrics of a classification report from machine learning and deep learning.

* **Precision:** Precision is the ability of a classifier not to label an instance positive that is actually negative. For each class it is defined as the ratio of true positives to the sum of true and false positives.

 …………… (3.4)

* **Recall:** Recall is the ability of a classifier to find all positive instances. For each class it is defined as the ratio of true positives to the sum of true positives and false negatives.

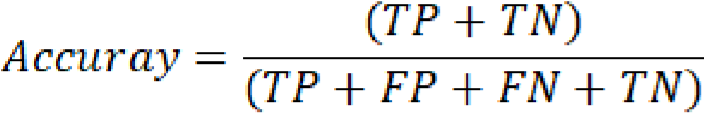
……………(3.5)

* **F1 Score:** The F1 score is a weighted harmonic mean of precision and recall such that the best score is 1.0 and the worst is 0.0. Generally speaking, F1 scores are lower than accuracy measures as they embed precision and recall into their computation. As a rule of thumb, the weighted average of F1 should be used to compare classifier models, not global accuracy.



……….. (3.6)

* **Support:** Number of actual occurrence of the class in the specified dataset.
* **Accuracy:** From accuracy we can calculate macro avg and weighted avg.



…………. (3.7)

**Macro Average:** Macro average is simply the average of the precision of the different class.

**Weighted Average:** Weighed average is the total number TP(true positive of all classes)/total number of objects in all classes.

Where TP, FP, FN, and TN represents the output measures as true positive, false positive, false negative, and true negative values for the training and validation images of the models.

**Accuracy Score comparison with Machine Learning approach**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Accuracy Score for Processed Dataset** | **Accuracy Score for Without Processed Dataset** | **Accuracy Score for Vgg16 feature extraction with Processed Dataset** | **Accuracy Score for Vgg16 feature extraction Without Processed Dataset** |
| **Decision Tree Classifier** | 0.68 | 0.65 | 0.61 | 0.63 |
| **K-Nearest Neighbors Classifier** | 0.70 | 0.65 | 0.69 | 0.72 |
| **Logistic Regression Classifier** | 0.75 | 0.71 | 0.79 | 0.82 |
| **Naive Bayes Classifier** | 0.58 | 0.60 | 0.69 | 0.69 |
| **Random Forest Classifier** | 0.76 | 0.73 | 0.80 | 0.83 |
| **Support Vector Machine Classifier** | 0.77 | 0.73 | 0.80 |  |

Table 2: Accuracy Score comparison with Machine Learning approach

From the above table we can see accuracy score for vgg16 feature extraction without processed dataset is best.

Figure 3.49: Accuracy Score comparison with Machine Learning approach

### Chapter 4

## Results and Discussion

#### 4.1 Matlab Result

The simulation results and classification of few CT images using our proposed method are included in Table 3.

| Image | ***Nodule*** | | ***Classifi-cation*** | ***Final Class*** |
| --- | --- | --- | --- | --- |
| Image 12 |  | Nodule 1 | Malignant | Malig-nant |
| Image 37 |  | Nodule 1 | Malignant | Malig-nant |
| Image 86 |  | Nodule 1 | Malignant | Malig-nant |
| Image 93 |  | Nodule 1 | Benign | Benign |
| Image 95 |  | Nodule 1  Nodule 2 | Malignant  Malignant | Malig-nant |
| Image 124 |  | Nodule 1 | Malignant | Malig-nant |
| Image  151 |  | Nodule 1 | Malignant | Malig-nant |

Table 3: nodules classification by proposed method

**BENIGN CASE CLASSIFICATION RESULTS**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Processed Dataset** | | | **Without Processed Dataset** | | | **Vgg16 Feature Extraction with Processed Dataset** | | | **Vgg16 Feature Extraction Without Processed Dataset** | | |
|  | **precision** | **recall** | **f1-score** | **precision** | **recall** | **f1-score** | **precision** | **recall** | **f1-score** | **precision** | **recall** | **f1-score** |
| **Decision Tree** | 0.42 | 0.38 | 0.40 | 0.29 | 0.31 | 0.30 | 0.15 | 0.38 | 0.22 | 0.27 | 0.39 | 0.32 |
| **K-Nearest Neighbors** | 0.18 | 0.35 | 0.24 | 0.22 | 0.28 | 0.25 | 0.17 | 0.38 | 0.24 | 0.15 | 0.22 | 0.18 |
| **Logistic Regression** | 0.39 | 0.73 | 0.51 | 0.29 | 0.39 | 0.33 | 0.40 | 0.69 | 0.51 | 0.37 | 0.53 | 0.43 |
| **Naive Bayes** | 0.21 | 0.42 | 0.28 | 0.07 | 0.14 | 0.10 | 1.00 | 0.12 | 0.21 | 0.17 | 0.03 | 0.05 |
| **Random Forest** | 0.48 | 0.62 | 0.54 | 0.33 | 0.14 | 0.20 | 0.47 | 0.31 | 0.37 | 0.27 | 0.19 | 0.23 |
| **Support Vector Machine** | 0.37 | 0.62 | 0.46 | 0.21 | 0.22 | 0.21 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |

Table 4: benign case classification results

From the above table we can see from machine learning algorithm svm cannot detect benign case because of poor amount of data. but doing vgg-16 extraction without processed data logistic regression model gives best accuracy.

Figure 4.50 Processed Dataset- benign case classification results

Figure 4.51 Without Processed Dataset- benign case classification results

From the above figure we can see for processed and unprocessed for both dataset logistic regression model is the best model to find all positive instances, which called Re-call.

Figure 4.52: Vgg16 Feature Extraction with Processed Dataset-BENIGN Case Classification Results

Figure 4.53: Vgg16 Feature Extraction without Processed Dataset-BENIGN Case Classification Results

From the above figure we can see for processed and unprocessed for both dataset logistic regression model is the best model in vgg-16 feature extraction to find all positive instances, which called Re-call.

**MALIGNANT CASE CLASSIFICATION RESULTS**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Processed Dataset** | | | | | | **Without Processed Dataset** | | | | | | **Vgg16 Feature Extraction with Processed Dataset** | | | | | | **Vgg16 Feature Extraction Without Processed Dataset** | | | | | |
|  | **precision** | | **recall** | | **f1-score** | | **precision** | | **recall** | | **f1-score** | | **precision** | | **recall** | | **f1-score** | | **precision** | | **recall** | | **f1-score** | |
| **Decision Tree Classifier** | 0.73 | | 0.74 | | 0.73 | | 0.75 | | 0.77 | | 0.76 | | 0.84 | | 0.70 | | 0.77 | | 0.80 | | 0.81 | | 0.80 | |
| **K-Nearest Neighbors Classifier** | 0.87 | | 0.78 | | 0.82 | | 0.80 | | 0.77 | | 0.78 | | 0.91 | | 0.79 | | 0.85 | | 0.98 | | 0.85 | | 0.91 | |
| **Logistic Regression Classifier** | 0.95 | | 0.74 | | 0.83 | | 0.84 | | 0.88 | | 0.86 | | 0.91 | | 0.84 | | 0.87 | | 0.95 | | 0.99 | | 0.97 | |
| **Naive Bayes Classifier** | 0.89 | | 0.60 | | 0.72 | | 0.91 | | 0.63 | | 0.75 | | 0.64 | | 0.99 | | 0.78 | | 0.67 | | 0.99 | | 0.80 | |
| **Random Forest Classifier** | | 0.83 | | 0.81 | | 0.82 | | 0.78 | | 0.86 | | 0.82 | | 0.84 | | 0.87 | | 0.85 | | 0.94 | | 0.99 | | 0.97 |
| **Support Vector Machine Classifier** | | 0.90 | | 0.83 | | 0.87 | | 0.87 | | 0.89 | | 0.88 | | 0.87 | | 0.84 | | 0.85 | | 0.98 | | 1.00 | | 0.99 |

Table 5: MALIGNANT CASE CLASSIFICATION RESULTS

From the above table we can see from machine learning algorithm svm(support vector machine) and logistic regression in all cases data gives best accuracy.

Figure 4.54: Machine Learning with Processed Dataset-Malignant Case Detection Result

Figure 4.55: Machine Learning with Processed Dataset-Malignant Case Detection Result

From above figure we can see in processed dataset (malignant cases) logistic regression gives best precision result and without processed dataset naïve bayes gives best precision result.

Figure 4.56: Vgg16 Feature Extraction with Processed Dataset-Malignant Case Detection Result

Figure 4.57: Vgg16 Feature Extraction without Processed Dataset-Malignant Case Detection Result

from above figure we can see in vgg-16 feature extraction (malignant case) in processed dataset Naïve Bayes gives best Re-call and in unprocessed dataset svm gives best classification report.

**NORMAL CASE CLASSIFICATION RESULTS**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Processed Dataset** | | | **Without Processed Dataset** | | | **Vgg16 Feature Extraction with Processed Dataset** | | | **Vgg16 Feature Extraction Without Processed Dataset** | | |
|  | **precision** | **recall** | **f1-score** | **precision** | **recall** | **f1-score** | **precision** | **recall** | **f1-score** | **precision** | **recall** | **f1-score** |
| **Decision Tree Classifier** | 0.66 | 0.66 | 0.66 | 0.62 | 0.60 | 0.61 | 0.60 | 0.54 | 0.57 | 0.54 | 0.46 | 0.49 |
| **K-Nearest Neighbors Classifier** | 0.69 | 0.67 | 0.68 | 0.61 | 0.60 | 0.60 | 0.67 | 0.60 | 0.63 | 0.65 | 0.68 | 0.66 |
| **Logistic Regression Classifier** | 0.69 | 0.77 | 0.73 | 0.69 | 0.58 | 0.63 | 0.77 | 0.74 | 0.75 | 0.82 | 0.67 | 0.74 |
| **Naive Bayes Classifier** | 0.48 | 0.60 | 0.53 | 0.59 | 0.68 | 0.63 | 0.93 | 0.40 | 0.56 | 0.79 | 0.47 | 0.59 |
| **Random Forest Classifier** | 0.72 | 0.71 | 0.71 | 0.61 | 0.62 | 0.62 | 0.78 | 0.79 | 0.79 | 0.76 | 0.78 | 0.77 |
| **Support Vector Machine Classifier** | 0.75 | 0.73 | 0.74 | 0.69 | 0.66 | 0.67 | 0.74 | 0.92 | 0.82 | 0.77 | 0.97 | 0.86 |

Table 6: NORMAL CASE CLASSIFICATION RESULTS

From table we can see in normal case svm and logistic regression gives best result.

Figure 4.58: Processed Dataset-Normal Case Classification Result

Figure 4.59: Without Processed Dataset-Normal Case Classification Result

From figure we can see svm and logistic regression gives best result.

Figure 4.60: Vgg16 Feature Extraction with Processed Dataset-Normal Case Classification Result

Figure 4.61: Vgg16 Feature Extraction without Processed Dataset-Normal Case Classification Result

From above figure we can see in normal case vgg-16 feature extraction (processed and unprocessed dataset) svm gives best result.

**Deep Learning Approach**

**Deep learning (cnn) Approach Accuracy Table**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Training Loss** | **Validation Loss** | **Training Accuracy** | **Validation Accuracy** | **Accuracy Score** |
| **CNN with Processed Dataset** | 2.5199e-04 | 0.1237 | 1.0000 | 0.9742 | 0.75 |
| **CNN without Processed Dataset** | 4.6185e-04 | 0.0019 | 1.0000 | 1.0000 | 0.71 |
| **Vgg16 feature extraction with Processed Dataset** | 0.0030 | 0.1933 | 1.0000 | 0.9548 | 0.80 |
| **Vgg16 feature extraction without Processed Dataset** | 0.0837 | 0.2022 | 0.9707 | 0.9351 | 0.84 |

Table 7: Deep learning (CNN) Approach Accuracy Table

From table we can see CNN with processed dataset training loss is less and vgg-16 feature extraction without processed dataset validation loss is low.

Figure 4.62: Deep learning (CNN) Approach Accuracy Graph

**Accuracy Curve and Loss Curve for CNN model on Processed Dataset**

Below, Fig. 4.63 and Fig. 4.64 shows the plot of model accuracy vs. epoch and model loss vs. epoch for training and validation images on processed data.

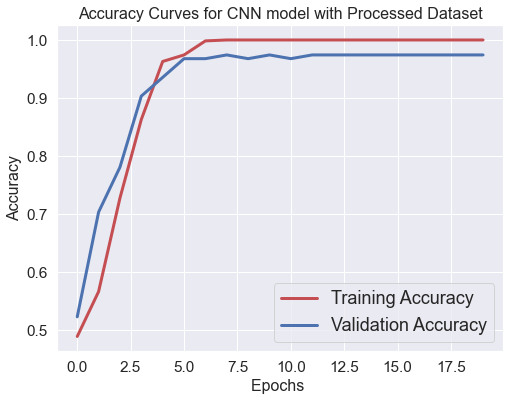
****

Figure 4.63: Plot of Model Accuracy vs. Epoch for Training and Validation Images

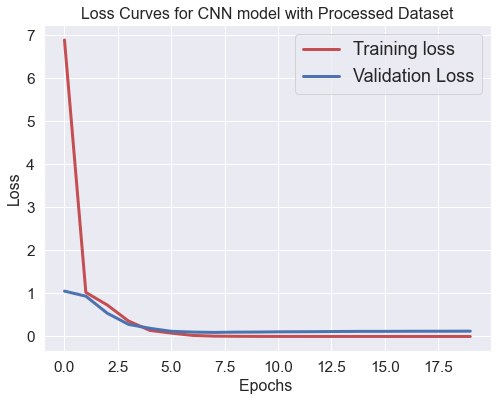
****

Figure 4.64: Plot of Model Loss vs. Epoch for Training and Validation Images

**Accuracy Curve and Loss Curve for CNN model on without Processed Dataset**

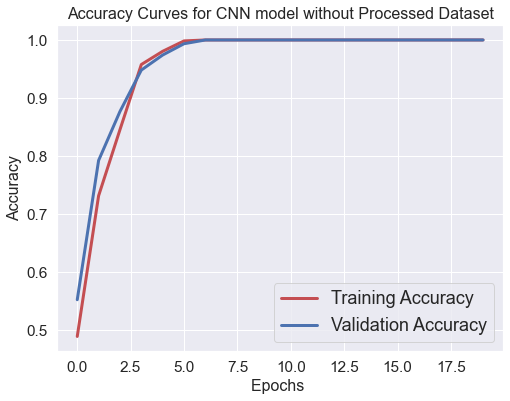
****Below, Fig. 4.65 and Fig. 4.66 shows the plot of model accuracy vs. epoch and model loss vs. epoch for training and validation images on processed data.

Figure 4.65: Accuracy Curve for CNN model on without Processed Dataset

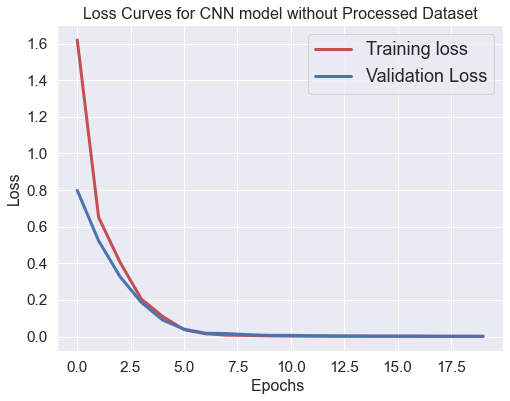
****

Figure 4.66: Loss for CNN model on without Processed Dataset

**Accuracy Curve and Loss Curve for VGG16 Feature extraction with processed Dataset**

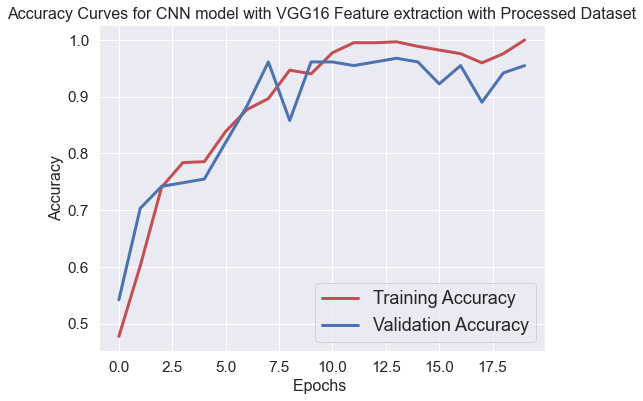
****

Figure 4.67: Accuracy Curve for VGG16 Feature extraction with processed Dataset

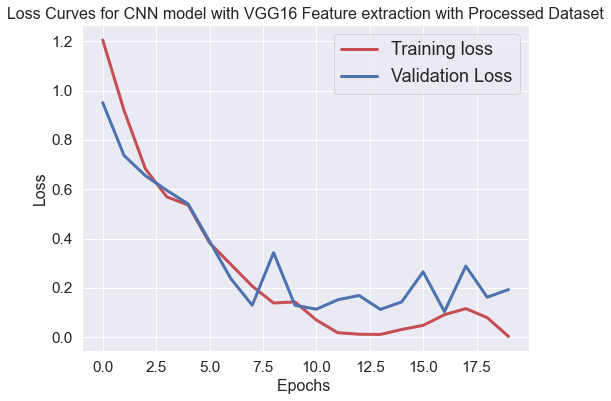
****

Figure 4.68: Loss Curve for VGG16 Feature extraction with processed Dataset

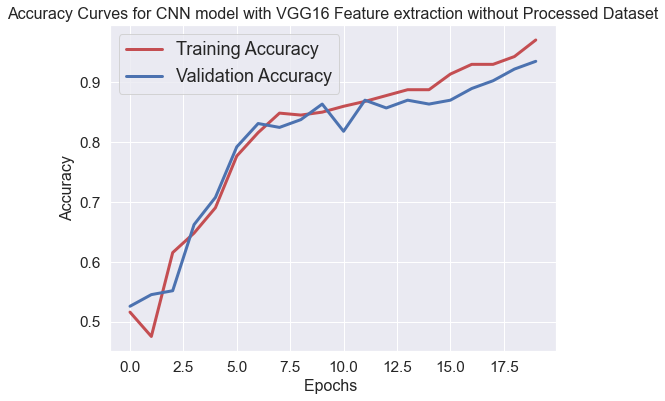
**Accuracy Curve and Loss Curve for VGG16 Feature extraction without processed Dataset**

Figure 4.69: Accuracy Curve for VGG16 Feature extraction without processed Dataset

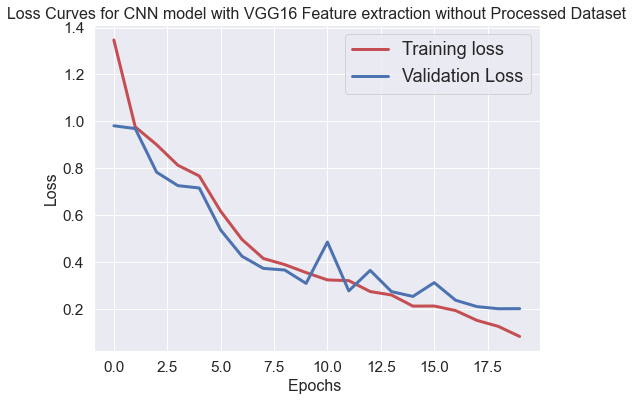
****

Figure 4.70: Loss Curve for VGG16 Feature extraction without processed Dataset

## Deep learning (CNN) Classification Results

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **CNN with Processed Dataset** | | | **CNN without Processed Dataset** | | | **Vgg16 feature Extraction with Processed Dataset** | | | **Vgg16 feature Extraction without Processed Dataset** | | |
|  | **precision** | **recall** | **f1-score** | **precision** | **recall** | **f1-score** | **precision** | **recall** | **f1-score** | **precision** | **recall** | **f1-score** |
| **BENIGN CASE CLASSIFICATION RESULTS** | 0.21 | 0.46 | 0.29 | 0.39 | 0.31 | 0.34 | 0.26 | 0.46 | 0.33 | 0.29 | 0.14 | 0.19 |
| **MALIGNANT CASE CLASSIFICATION RESULTS** | 0.94 | 0.88 | 0.91 | 0.76 | 0.92 | 0.83 | 0.89 | 0.96 | 0.92 | 0.96 | 0.96 | 0.96 |
| **NORMAL CASE CLASSIFICATION RESULTS** | 0.76 | 0.64 | 0.70 | 0.69 | 0.53 | 0.60 | 0.88 | 0.63 | 0.74 | 0.76 | 0.87 | 0.81 |

Table 8: Deep learning (CNN) Classification Results

#### 4.3 Classification Report with chart

The table shows the precision, recall, support and f-score for the different histopathology image categories. The formula to calculate the given metrics is explained show in analysis section.

Figure 4.71: CNN with Processed dataset

Figure 4.72: CNN without Processed Dataset

Figure 73: Vgg16 Feature Extraction with Processed Dataset

Figure 4.74: Vgg16 Feature Extraction without Processed Dataset

## Chapter 5

### Conclusion and Recommendation

5.1 CONCLUSION

A respiratory organ tumor organization has been designed and developed. This work presents a replacement approach to the automatic classification of respiratory organ tumors supported texture features, that separate respiratory organ tumor pictures from healthy tissues in computerized tomography images. The feature image used for the tumor classification consists of computed tomography pictures. the appliance of the planned methodology for pursuit tumor is incontestable to help pathologists distinguish its kind of respiratory organ tumor. A classification with Associate in Nursing accuracy of eighty eight, and eighty two, has been obtained by support vector and logistic regression with vgg16 extraction without processed data by machine and convolutional neural networks.

In earlier times, the doctor has to do multiple tests in order to detect whether a given patient has lung cancer or not . But this was a very time consuming process. In a diagnosis sometimes a patient has to undergo unnecessary check-ups or different tests to identify the disease of lung cancer. To minimize the process time and unnecessary check-ups there needs to be a preliminary test in which both the patient and the doctor will be notified with the possibilities of lung cancer. Nowadays the machine learning algorithms plays an important role in the prediction and classification of medical data. Logistic Regression, SVM, decision tree and Naïve Bayes, KNN, Random Forest ,CNN are the machine learning algorithms used for this comparative study. A comparative analysis of accuracy rates of each classifier are presented. The predictive performance of classifiers are compared quantitatively. In the performance chart, different results are produced for each classifier on the lung cancer dataset. Looking at the correct classification (CA) and other metrics; the best result is given by the support vector machine algorithm and logistic regression. SVM algorithm used high dimension to classify the observation so it’s performance is the best.

When using this trained network for the prediction purpose of our pipeline, first we pre-processed data by image processing method in matlab. Then the dataset was divided into two set one is for test another is for train. Then each of the training sets were sent through different kind of machine learning algorithms. It can be seen that all these algorithms ( SVM, DT, k-NN ,CNN, logistic regression ) SVM and logistic gives best accuracy. Then we again extract feature with vgg-16 and sent through machine learning algorithms where it gives significant result from support vector machine.

#### **5.2 Limitations**

From a procedure perspective, bioinformatics datasets, like the one studied here, challenge the boundaries of the progressive applied machine learning ways in many ways: little sample size will increase the danger of overfitting model parameters to the information. ripping the information set more to perform parameter-optimizing cross validation creates the danger of ending up with sample subsets that area unit too little to be helpful. High rates of missing values compromise the educational ability of machine learning algorithms, whereas terribly giant variable-to-sample ratios impede even ways that area unit sometimes sturdy to terribly high spatial property, as was incontestable in our experiments.

#### 5.3 Recommendations

The following suggestions could be included for same analysis works in future.

* The dataset has been chosen for conducting research was low resolution so this is important to collect high resolution image for this type of research.
* Raw images cannot identify the accuracy so correction is needed before applying any indices.
* This study used machine and deep learning algorithms for lung tumor classification but the further research can apply more accurate classification to detect lung tumor class.
* This research used k-NN, CNN, DT, LR, NB algorithms to identify malignant, normal and benign tumor but further analysis can detect more features this type of research.
* There is a mixture of lung, malignant and normal tumor which can be identify by doctor for better result to conduct this result.
* This analysis gives case research for some country which is probably not represent of different country. Future analysis may focus on applying our strategies to different country of the world and evaluating whether our strategies might be properly generalized.
* Important analysis of lung tumor growth is recommended.

# References

|  |  |
| --- | --- |
| [1] | American Cancer Society, Lung Cancer Alliance, American Psychosocial oncology society, "Top masters in health care," [Online]. Available: https://www.topmastersinhealthcare.com/lung-cancer/. |
| [2] | WHO, "WHO REPORT ON CANCER," WHO, 2020. |
| [3] | International Agency for Reseaerch on Cancer, "GCO," GLOBOCAN2020, 2020. [Online]. Available: https://gco.iarc.fr/today/data/factsheets/cancers/15-Lung-fact-sheet.pdf. |
| [4] | A. D. Moffy Vas, "Lung cancer detection system using lung CT image," in *IEEE*, 17-18 Aug. 2017. |
| [5] | S. A. A. H. Janee Alam, "Multi-Stage Lung Cancer Detection and Prediction Using Multi-class SVM Classifier," in *2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2)*, 8-9 Feb. 2018. |
| [6] | J. Brownlee, "Machine Learning MIstry," 17 April 2020. [Online]. Available: https://machinelearningmastery.com/convolutional-layers-for-deep-learning-neural-networks/. |
| [7] | Wikimedia Foundation, Inc.,, "Wikipedia," 13 12 2020. [Online]. Available: https://en.wikipedia.org/wiki/AlexNet. |
| [8] | M. u. Hassan, "neurohive," 20 11 2020. [Online]. Available: https://neurohive.io/en/popular-networks/vgg16/. |
| [9] | "The MathWorks, Inc.," [Online]. Available: https://www.mathworks.com/help/deeplearning/ref/googlenet.html;jsessionid=4105508e70451735c49e1499f9da. |
| [10] | Y.-c. H. C.-J. L. D. H. M. D. H. J. O. G. M. H. V. Jue Jiang, "Multiple Resolution Residually Connected Feature Streams For Automatic Lung Tumor Segmentation From CT Images," in *IEEE Transactions on Medical Imaging*, 2018. |
| [11] | N. K. a. U. Bagci, "Semi-Supervised Multi-Task Learning for Lung Cancer Diagnosis," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2018. |
| [12] | M. B. Vesna Zeljkovic, "Automatic Detection Of Abnormalities In Lung Radiographs Caused By Planocellular Lung Cancer," in *2011 1st Middle East Conference on Biomedical Engineering*, 2011. |
| [13] | M. Sakamoto, S. Sora and F. Kuroki, "Flat-types of Electrodes to Estimate Lung Cancer Position Using Resonant Circuit Technology," in *2020 IEEE International Symposium on Radio-Frequency Integration Technology (RFIT)*, 2020. |
| [14] | B. M. C. M. A. F. S. M. P. M. Lauren E.CiprianoBSc, "Lung cancer treatment costs, including patient responsibility, by disease stage and treatment modality, 1992 to 2003," *Elsevier,* 2011. |
| [15] | K. J. Danjuma, "Performance Evaluation of Machine Learning Algorithms in Post-operative Life Expectancy in the Lung Cancer Patients," *IJCSI International Journal of Computer Science ,* vol. 12, no. 2, 2015. |
| [16] | M. Kaur and R. Mittal, "Survey of Intelligent Methods for Brain Tumor Detection," *International Journal of Computer Science Issues (IJCSI,* vol. 11, no. 5, 2014. |
| [17] | T. T. Z Karhan, "Lung Cancer Detection and Classification with Classification Algorithms," *IOSR Journal of Computer Engineering (IOSR-JCE),* vol. 18, no. 6, 2016. |
| [18] | R. Ada, "A Study of Detection of Lung Cancer Using Data Mining Classification Techniques," *International Journal of Advanced Research in Computer Science and Software Engineering,* vol. 3, no. 3, 2013. |
| [19] | S. S. S. Anita Chaudhary, "Lung cancer detection on CT images using image processing," *Computing sciences 2012 international conference,IEEE,* 2012. |
| [20] | A. T. Z. A.-M. M. M. J. P. Mohamed Shakeel, "Automatic detection of lung cancer from biomedical data set using discrete AdaBoost optimized ensemble learning generalized neural networks," *Springer,* 2019. |
| [21] | S. Hussein, P. Kandel, C. W. Bolan, M. B. Wallace and U. Bagci, "Lung and Pancreatic Tumor Characterization in the Deep Learning Era: Novel Supervised and Unsupervised Learning Approaches," *IEEE Transactions on Medical Imaging,* vol. 38, no. 8, 2019. |
| [22] | A. A. &. A. Srinivasan, "Deep learning for lung Cancer detection and classification," *springer,* 2020. |
| [23] | J. A. M. S.-G. &. C. J. Sidey-Gibbons, "Machine learning in medicine: a practical introduction," *BMC Medical Research Methodology ,* vol. 19, no. 1, pp. 1-18, 2019. |
| [24] | P. BrinthaThereseDr, "Detection of Cancer in Lung with K-NN Classification Using Genetic Algorithm," *Procedia Materials Science,* vol. 10, pp. 433-440, 2015. |
| [25] | K. H. ,. T. a. C. P. May Phu Paing, "Automatic detection and staging of lung tumors using locational features and double-staged classifications," *Applied Science,* vol. 9, no. 11, p. 2329, 2019. |
| [26] | G. A. P. S. &. P. K. Gupta, "Performance analysis of various machine learning-based approaches for detection and classification of lung cancer in humans," *Natural Computing and Applications,* vol. 31, no. 10, pp. 6863-6877, 2018. |
| [27] | M. Vas and A. Dessai, ". Lung cancer detection system using lung CT image processing.," in *2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA)*, 2017. |
| [28] | G. Jakimovski and D. Davcev, "Lung cancer medical image recognition using Deep Neural Networks," in *2018 Thirteenth International Conference on Digital Information Management (ICDIM)*, 2018. |
| [29] | N. S. Nadkarni and S. Borkar, "Detection of lung cancer in CT Images using image processing.," in *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, 2019. |
| [30] | M. Rumman, A. N. Tasneem, S. Farzana, M. I. Pavel and M. A. Alam, "Early detection of Parkinson’s disease using image processing and artificial neural network," in *2018 Joint 7th International Conference on Informatics, Electronics & Vision (ICIEV) and 2018 2nd International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*, 2018. |
| [31] | P. S. R. M. Punithavathy K, " Performance evaluation of machine learning techniques in lung cancer classification from PET/CT images.," *FME Transactions,* vol. 17, no. 3, pp. 418-423, 2019. |
| [32] | "MathWorks," [Online]. Available: https://www.mathworks.com/help/images/ref/medfilt2.html#bupa05x-1\_seealso.. [Accessed 24 April 2021]. |
| [33] | "MathWorks," [Online]. Available: https://www.mathworks.com/help/images/ref/imadjust.html. [Accessed 23 March 2021]. |
| [34] | "MathWorks," [Online]. Available: https://www.mathworks.com/discovery/image-thresholding.html.. [Accessed 12 MArch 2021]. |
| [35] | N. OTSU, "A Tlreshold Selection Method," *IEEE TRANSACTIONS ON SYSTREMS, MAN, AND CYBERNETICS, V,* vol. 9, no. 1, 1979. |
| [36] | "MathWorks," [Online]. Available: https://www.mathworks.com/help/images/ref/graythresh.html#bviewj6. [Accessed 15 march 2021]. |
| [37] | H. KArim, "Kaggle," 24 May 2020. [Online]. Available: https://www.kaggle.com/hamdallak/the-iqothnccd-lung-cancer-dataset. |
| [38] | R. P.R., R. A. Nair and V. G., "A Comparative Study of Lung Cancer Detection using Machine Learning Algorithms," in *2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, 2019. |