# CT Scan Image Processing for Lung Cancer Detection and Classification using Different Machine Learning and Deep Learning Methods



**Undergraduate Thesis**

**Dept. of Electrical and Electronic Engineering**

**Ahsanullah University of Science and Technology**

The thesis is presented in partial fulfillment of the criteria for the degree of Bachelor of Science(B.Sc.) in Electrical and Electronic Engineering

**Supervised by**

**Dr. Umma Hany(P.hD)**

**Associate Professor**

Dept. of Electrical and Electronic Engineering

Ahsanullah University of Science and Technology

**Submitted by**

Nusraat Nawreen [ID: 160205091]

Tahmina Islam [ID : 160205092]

Nafe Muhtasim Hye [ID :160205095]

Abdullah Al Mamun [ID : 160205103]

# APPROVAL

The thesis titled “**CT Scan Image Processing for Lung Cancer Detection and Classification using Different Machine Learning and Deep Learning Methods**” submitted by Nusraat Nawreen, Tahmina Islam, Nafe Muhtasim Hye, Abdullah Al Mamun,Session: Spring 2020 has been accepted as satisfactory in partial fulfillment of the requirement for the Degree of BACHELOR OF SCIENCE IN ELECTRICAL AND ELECTRONIC ENGINEERING.

**Supervisor**

Dr.Umma Hany

Assosiate Professor

Department of Electrical and Electronic Engineering

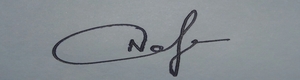
# DECLARATION

This is to certify that the thesis titled “**CT Scan Image Processing for Lung Cancer Detection and Classification using Different Machine Learning and Deep Learning Methods**” is the result of our study in partial fulfillment of the B.sc Engineering degree under the supervision of Dr. Umma Hany, Assosiate Professor,Department of Electrical and Electronic Engineering(EEE),Ahsanullah University of Science and Technology (Aust), and it has not been submitted elsewhere for any other degree or diploma.

**Student Names and Signature**

Nusraat Nawreen [16-02-05-091]

Tahmina Islam [ID : 16-02-05-092]



Nafe Muhtasim Hye [ID :160205095]

Abdullah Al Mamun [ID : 160205103]

# 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**. 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 have acquired significant accuracy in locating lung tumor applying 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 want 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

july, 2021

Thankfully

Nusraat Nawreen

Tahmina Islam

Nafee Muhtasim

Abdullah Al Mamun

# 

Contents

[ABSTRACT 2](#_Toc76817322)

[ACKNOWLEDGEMENT 3](#_Toc76817323)

[Abbreviations and Acronyms 7](#_Toc76817324)

[Chapter 1 8](#_Toc76817325)

[Introduction 8](#_Toc76817326)

[1.1 Lung Cancer 8](#_Toc76817327)

[1.2 Objective of the Research 9](#_Toc76817328)

[1.3 Motivation of this thesis 10](#_Toc76817329)

[1.4 Structure of the Thesis 10](#_Toc76817330)

[Chapter 2 12](#_Toc76817331)

[Literature Review 12](#_Toc76817332)

[2.1 Literature Review 12](#_Toc76817333)

[Chapter 3 15](#_Toc76817334)

[3.1 Data Collection 15](#_Toc76817335)

[3.2 Image Preprocessing: 15](#_Toc76817336)

[3.2.1: Image Resizing 15](#_Toc76817337)

[3.2.2: Image Smoothing 15](#_Toc76817338)

[3.2.3 Image Enhancement 16](#_Toc76817339)

[3.3 Histogram Plots 16](#_Toc76817340)

[3.4 Image Segmentation 16](#_Toc76817341)

[3.5 Thresholding: 16](#_Toc76817342)

[3.6 Edge Detection 17](#_Toc76817343)

[3.7 Feature Extraction 17](#_Toc76817344)

[3.8 Simulation Result: 18](#_Toc76817345)

[Lung tumor classification using machine learning method 20](#_Toc76817346)

[4.1 Machine learning: 20](#_Toc76817347)

[4.2 Classification of lung tumor using deep learning and machine learning algorithms: 20](#_Toc76817348)

[4.3 Vgg16 (Transfer learning) 20](#_Toc76817349)

[4.4 K-Nearest Neighbors Classifier 21](#_Toc76817350)

[4.5 SVM (support vector machine) 22](#_Toc76817351)

[4.6 Decision Tree 22](#_Toc76817352)

[4.7 Naive Bayes 23](#_Toc76817353)

[4.8 Logistic Regression: 24](#_Toc76817354)

[4.9 Random forest classifier: 25](#_Toc76817355)

[1.Preprocessing Phase: 25](#_Toc76817356)

[2.Prediction Phase: 25](#_Toc76817357)

[System flow architecture 26](#_Toc76817358)

[Support Vector Machine on processed Dataset 26](#_Toc76817359)

[Logistic Regression Classifier on processed Dataset 28](#_Toc76817360)

[Random Forest Classifier on processed Dataset 29](#_Toc76817361)

[Unprocessed Dataset 30](#_Toc76817362)

[Support Vector Machine on Unprocessed Dataset 30](#_Toc76817363)

[Logistic Regression on unprocessed Dataset 31](#_Toc76817364)

[Random Forest Classifier on unprocessed Dataset 32](#_Toc76817365)

[Vgg16 feature extraction with Processed Dataset 33](#_Toc76817366)

[Support Vector Machine on Vgg16 feature extraction with Processed Dataset 33](#_Toc76817367)

[Support Vector Machine on Vgg16 feature extraction without Processed Dataset 37](#_Toc76817368)

[Logistic Regression on Vgg16 feature extraction without Processed Dataset 38](#_Toc76817369)

[Random Forest Classifier on Vgg16 feature extraction without Processed Dataset 39](#_Toc76817370)

[Classification Report: 40](#_Toc76817371)

[Benign Classification Result 41](#_Toc76817372)

[5.1 Deep learning 50](#_Toc76817373)

[5.2 CNN 50](#_Toc76817374)

[Input Layer 51](#_Toc76817375)

[Convolutional 2D 51](#_Toc76817376)

[Relu 51](#_Toc76817377)

[Maxpooling 51](#_Toc76817378)

[Dense Layer 51](#_Toc76817379)

[Output 51](#_Toc76817380)

[CNN with Processed Dataset 52](#_Toc76817381)

[CNN without Processed Dataset 54](#_Toc76817382)

[Deep learning (CNN) Classification Results 55](#_Toc76817383)

[Classification Report with chart 55](#_Toc76817384)

[Deep Learning Approach 59](#_Toc76817385)

[Accuracy Curve and Loss Curve for CNN model on Processed Dataset 61](#_Toc76817386)

[Accuracy Curve and Loss Curve for CNN model on without Processed Dataset 62](#_Toc76817387)

[Accuracy Curve and Loss Curve for VGG16 Feature extraction with processed Dataset 63](#_Toc76817388)

[Chapter 6 64](#_Toc76817389)

[Monitoring and Analysis of system model 64](#_Toc76817390)

[Accuracy Score comparison with Machine Learning approach 64](#_Toc76817391)

[Chapter 7 67](#_Toc76817392)

[Conclusion and Recommendation 67](#_Toc76817393)

[7.1 CONCLUSION 67](#_Toc76817394)

[7.2 Limitations 68](#_Toc76817395)

[7.3 Future works 68](#_Toc76817396)

[References 70](#_Toc76817397)

**List of Tables:**

[Table 1. 1: Lung Cancer in US 10](#_Toc76822381)

[Table 4. 1:benign case classification results 41](#_Toc76822395)

[Table 4. 2: Malignant Case Classification Results 44](#_Toc76822396)

[Table 4. 3: Normal Case Classification Results 47](#_Toc76822397)

[Table 5. 1: Deep learning (CNN) Classification Results 54](#_Toc76822476)

[Table 5. 2 **:** Deep learning (CNN) Approach Accuracy Table 58](#_Toc76822477)

[Table 6. 1: Accuracy Score comparison with Machine Learning approach 63](#_Toc76822490)

**List of Figures:**

[Figure 3. 1: Two of the Lung disease CT images and their Histogram 18](#_Toc76822592)

[Figure 3. 2: Cropped image and the median filtered image of the CT images 18](#_Toc76822593)

[Figure 3. 3: The Enhanced Images and their Histograms 19](#_Toc76822594)

[Figure 3. 4: Edges of the Binary images and the Edge subtracted Binary images 19](#_Toc76822595)

[Figure 3. 5: ROI of the Lung Tumor and their Classification 19](#_Toc76822596)

[Figure 4. 1: Network Structure 23](#_Toc76822603)

[Figure 4. 2: Support Vector Machine 24](#_Toc76822604)

[Figure 4. 3: Example X data and Probability-Class Relations 25](#_Toc76822605)

[Figure 4. 4: Logistic Regression to distinguish two classes 26](#_Toc76822606)

[Figure 4. 5: System Flow Architecture 28](#_Toc76822607)

[Figure 4. 6: Example of predicted Images of the Test Data using SVM with Processed Dataset 28](#_Toc76822608)

[Figure 4. 7: Example of Mispredicted Images of the Test Data using CNN without Processed Dataset 29](#_Toc76822609)

[Figure 4. 8: Example of predicted Images of the Test Data using Logistic Regression Classifier on Processed Dataset 30](#_Toc76822610)

[Figure 4. 9: Example of Mispredicted Images of the Test Data using Logistic Regression Classifier on Processed Dataset. 30](#_Toc76822611)

[Figure 4. 10: Example of predicted Images of the Test Data using Random Forest Classifier on processed Dataset 31](#_Toc76822612)

[Figure 4. 11: Example of Mispredicted Images of the Test Data using Random Forest Classifier on processed Dataset 31](#_Toc76822613)

[Figure 4. 12: Example of predicted Images of the Test Data using Support Vector Machine on Unprocessed Dataset 32](#_Toc76822614)

[Figure 4. 13: Example of Mispredicted Images of the Test Data using Support Vector Machine on Unprocessed Dataset 33](#_Toc76822615)

[Figure 4. 14: Example of predicted Images of the Test Data using Logistic Regression on unprocessed Dataset. 33](#_Toc76822616)

[Figure 4. 15: Example of Mispredicted Images of the Test Data using Logistic Regression on unprocessed Dataset. 34](#_Toc76822617)

[Figure 4. 16: Example of predicted Images of the Test Data using Random Forest Classifier on unprocessed Dataset 34](#_Toc76822618)

[Figure 4. 17: Example of Mispredicted Images of the Test Data using Random Forest Classifier on unprocessed Dataset. 35](#_Toc76822619)

[Figure 4. 18: Example of predicted Images of the Test Data using Support Vector Machine on Vgg16 feature extraction with Processed Dataset 36](#_Toc76822620)

[Figure 4. 19: Example of mispredicted Images of the Test Data using Support Vector Machine on Vgg16 feature extraction with Processed Dataset 36](#_Toc76822621)

[Figure 4. 20: Example of predicted Images of the Test Data using Random Forest Classifier on Vgg16 feature extraction with Processed Dataset 37](#_Toc76822622)

[Figure 4. 21: Example of mispredicted Images of the Test Data using Random Forest Classifier on Vgg16 feature extraction with Processed Dataset. 37](#_Toc76822623)

[Figure 4. 22: Example of predicted Images of the Test Data using Logistic Regression on Vgg16 feature extraction with Processed Dataset. 38](#_Toc76822624)

[Figure 4. 23: Example of Mispredicted Images of the Test Data using Logistic Regression on Vgg16 feature extraction with Processed Dataset. 38](#_Toc76822625)

[Figure 4. 24: Example of predicted Images of the Test Data using Support Vector Machine on Vgg16 feature extraction without Processed Dataset. 39](#_Toc76822626)

[Figure 4. 25: Example of Mispredicted Images of the Test Data using Support Vector Machine on Vgg16 feature extraction without Processed Dataset. 39](#_Toc76822627)

[Figure 4. 26: Example of predicted Images of the Test Data using Logistic Regression on Vgg16 feature extraction without Processed Dataset 40](#_Toc76822628)

[Figure 4. 27: Example of Mispredicted Images of the Test Data using Logistic Regression on Vgg16 feature extraction without Processed Dataset 41](#_Toc76822629)

[Figure 4. 28: Example of predicted Images of the Test Data using Random Forest Classifier on Vgg16 feature extraction without Processed Dataset. 41](#_Toc76822630)

[Figure 4. 29: Example of Mispredicted Images of the Test Data using Random Forest Classifier on Vgg16 feature extraction without Processed Dataset 41](#_Toc76822631)

[Figure 4. 30: Processed Dataset- benign case classification results 44](#_Toc76822632)

[Figure 4. 31: Without Processed Dataset- benign case classification results 44](#_Toc76822633)

[Figure 4. 32: Vgg16 Feature Extraction with Processed Dataset-BENIGN Case Classification Results. 45](#_Toc76822634)

[Figure 4. 33: Vgg16 Feature Extraction without Processed Dataset-BENIGN Case Classification Results. 45](#_Toc76822635)

[Figure 4. 34 : Machine Learning with Processed Dataset-Malignant Case Detection Result 47](#_Toc76822636)

[Figure 4. 35: Machine Learning with Processed Dataset-Malignant Case Detection Result 47](#_Toc76822637)

[Figure 4. 36: Vgg16 Feature Extraction with Processed Dataset-Malignant Case Detection Result 48](#_Toc76822638)

[Figure 4. 37: Vgg16 Feature Extraction without Processed Dataset-Malignant Case Detection Result. 48](#_Toc76822639)

[Figure 4. 38: Processed Dataset-Normal Case Classification Result 50](#_Toc76822640)

[Figure 4. 39: Without Processed Dataset-Normal Case Classification Result. 50](#_Toc76822641)

[Figure 4. 40: Vgg16 Feature Extraction with Processed Dataset-Normal Case Classification Result. 51](#_Toc76822642)

[Figure 4. 41: Vgg16 Feature Extraction without Processed Dataset-Normal Case Classification Result. 51](#_Toc76822643)

[Figure 5. 1: CNN Architecture 54](#_Toc76822644)

[Figure 5. 2: Example of Processed Image of the Test Dataset Using CNN with Processed Dataset 55](#_Toc76822645)

[Figure 5. 3: Example of mispredicted Images of the Test Data using CNN with Processed Dataset. 56](#_Toc76822646)

[Figure 5. 4: Examples of Predicted Image of the Test Dataset using CNN without Processed Dataset 56](#_Toc76822647)

[Figure 5. 5: Example of Mispredicted Images of the Test Data using CNN without Processed Dataset. 57](#_Toc76822648)

[Figure 5. 6: CNN with Processed dataset 58](#_Toc76822649)

[Figure 5. 7: CNN without Processed Dataset 59](#_Toc76822650)

[Figure 5. 8: Vgg16 Feature Extraction with Processed Dataset 60](#_Toc76822651)

[Figure 5. 9: Vgg16 Feature Extraction without Processed Dataset. 60](#_Toc76822652)

[Figure 5. 10: Deep learning (CNN) Approach Accuracy Graph 62](#_Toc76822653)

[Figure 5. 11: Plot of Model Accuracy vs. Epoch for Training and Validation Images. 63](#_Toc76822654)

[Figure 5. 12: Plot of Model Loss vs. Epoch for Training and Validation Images 63](#_Toc76822655)

[Figure 5. 13: Accuracy Curve for CNN model on without Processed Dataset 64](#_Toc76822656)

[Figure 5. 14: Loss for CNN model on without Processed Dataset 64](#_Toc76822657)

[Figure 5. 15: Accuracy Curve for VGG16 Feature extraction with processed Dataset 65](#_Toc76822658)

[Figure 5. 16: Loss Curve for VGG16 Feature extraction with processed Dataset 65](#_Toc76822659)

[Figure 6. 1: Accuracy Score comparison with Machine Learning approach 67](#_Toc76822660)

# 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

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

### 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.

Cancer cells tend to spread really fast due to blood streams and lymph fluid that is present in lung tissue. In general, due to normal lymph flow, cancer cells frequently migrate to the middle of the chest. As cancer cells migrate to other tissues, metastasis occurs. It is important that cancer be detected as early as possible as it tends to spread and is beyond curable in case of a larger spread. 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.

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.

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]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. 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.

### 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 classify the lung tumor cancer using CNN (deep learning).

To apply different types of machine learning algorithm to classify lung tumor.

To compare the accuracy between different types of algorithm.

### 1.3 Motivation of this thesis

* From January 2015 to December 2017, 5,887 people with lung cancer were admitted to the hospital.
* Lung was the leading site of cancers in men.24.7% men were admitted with lung cancer and 5.2% women.
* 77.2% of patients did not receive any kind of cancer treatment before attending NICRH
* One-third of cancer patients in Bangladesh are admitted to hospital.

Table 1. 1: Lung Cancer in US

|  |  |  |
| --- | --- | --- |
| **Lung & Bronchus Cancer in US** | **Male** | **Female** |
| Estimated  New Cases | 119,100 | 116,660 |
| Estimated New Deaths | 69,410 | 62,470 |

### 1.4 Structure of the Thesis

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

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

Second chapter is the literature review 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.

In the fourth chapter lung tumor classification using machine learning algorithms and results.

In fifth chapter we explained the lung tumor classification using deep learning algorithms.

Then in fourth chapter we showed the comparison and accuracy result of our analysis.

The final chapter is conclusion, future works and limitations.

# Chapter 2

## Literature Review

The detailed of literature reviews of this research has been stated in this chapter.

### 2.1 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. . 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].

Manages the forecast of post-usable future in cellular breakdown in the lungs patients utilizing prescient information mining calculations to analyze calculations, for example, Decision Tree, Naive Bayes and Artificial neural organization. A separated 10-crease cross-approval relative investigation was led on the above calculations and precision was determined for every classifier [14].

Paper manages similar investigation of characterization calculation for identification of Brain Tumor. Utilizing volumetric and area highlights generally speaking exactness rate was determined dependent on 2 characterization classes, for example, calculated relapse and Quadratic Discriminant and 3 arrangement classes like Linear SVM, Coarse Gaussian SVM, Cosine KNN and Complex and middle tree[15].

In this paper, various outcomes are delivered for every classifier on the cellular breakdown in the lungs dataset acquired. The classifiers, for example, KNN, SVM, NN and Logistic Regression were carried out and relating exactness rates were gotten. Backing Vector Machine has the most elevated exactness with 99.3%.The proposed strategy was applied to clinical dataset which assisted specialists with settling on more right choice [16]

Different division calculations were examined which incorporates Naïve Bayes, Hidden Markov Model and so on Legitimate clarification is given about how and why different division calculations are utilized in identification of Lung tumor [17].

In [18], 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 [19], 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 [20], 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 [21], 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 [22], 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 [23], 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 [24], 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 [25], 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 [26], 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 cancer cell. Hence they have decided to take the result as accurate if the certainty of cancer is 75%. In [27], 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 [28], 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 [29], 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.

# Chapter 3

#### Lung Tumor detection using CT scan images

Image processing is a technique for applying operations on an image in order to improve it or extract relevant information from it. In this chapter, image processing technique for lung tumor detection is discussed.

### 3.1 Data Collection

The dataset used for lung tumor detection is collected from The Cancer Imaging Archive (TCIA) and originally it contained 56 dicom images. For checking the feasibility the lung tumor detection methods we’ve augmented these 56 images into 168 images and applied the detection technique into it.

### 3.2 Image Preprocessing:

The main objective of image preprocessing is to eliminate unwanted distortions present in the image and enhance some useful features for further processing.

In the image pre-processing steps, we’ve done –

1. Image resizing
2. Image Smoothing
3. Image Enhancement

### 3.2.1: Image Resizing

Firstly the original image was grayscale and in 512x512 size. We’ve converted the image into 256 x 256 dimensions and adjusted it by cropping some extensions for better viewing.

### 3.2.2: Image Smoothing

In image smoothing, it is actually done to remove unwanted noise present in the image. As CT scanned images are prone to salt and pepper noise, we found the median filter pretty effective for removing such noise while preserving the edges. The median filter is a non-linear filter that uses zero padding concept so that the edges are unchanged while smoothing the other part of the image.

### 3.2.3 Image Enhancement

Image enhancement is the process of adjusting images so that the results are more suitable for displaying and further work. It produces a more suitable result than the original image for a specific application.

We use it to remove noise, sharpen or brighten an image to make it easier to identify the key features.

An image with good contrast has sharp difference between black and white. Here we’ve used the MATLAB tool named imadjust to adjust the enhancement of the image. This tool maps the intensity of the image into a new intensity valued image. By default, it sharpens the 1% bottom and top 1% of all the pixels.

### 3.3 Histogram Plots

The x-axis of the histogram represents the luminance value or intensity of the grey level and the y-axis of the histogram represents the number of pixel counts.

The result shows original image has more luminance values on the left side which means the original image was darker and after pre-processing of the image, the pixel are distributed in the entire grey level.

### 3.4 Image Segmentation

Image segmentation plays a vital role in detecting tumors. Image segmentation is typically used to locate objects and boundaries. By dividing an image into segments, we can process only the important segments of the image instead of processing the entire image.

We’ve implemented image segmentation to locate and segment out the lung tumor or our desired region of interest (ROI).

We’ve applied global thresholding using Otsu’s method and edge detection using Robert’s, followed by contrast adjustment to segment the lung nodule.

### 3.5 Thresholding:

In digital image processing, thresholding is the simplest method of segmenting the image. From a greyscale image, thresholding can be used to create a binary image. We use thresholding to compute one level so that when all the pixels in the input grayscale image with luminance is greater that the level, then it’s replaced by 1 or white and when all the pixels in the input greyscale image with luminance is less than the level, then it’s replaced with 0 or black. Here the level indicates the intensity level of the image.

Global thresholding using Otsu’s method has been performed here. In computer vision and image processing, Otsu’s method is used for automatic thresholding. The algorithm returns a single intensity threshold that separates pixels into two classes – foreground and background.

This threshold level is determined by minimizing intra class variance and maximizing the inter class variances. This method ignores the non-zero imaginary part of the image. The threshold value is normalized between 0 and 1.

### 3.6 Edge Detection

Edge detection is an image processing technique for finding the boundaries of the object within the image.

We have performed edge detection using Roberts’s method. Roberts’s method finds edges at those point where the gradient of the image is maximum.

After applying this, we have subtracted the edges from the binary image. We use contrast adjustment to adjust the intensity values of the edge subtracted binary image.

Finally, clear border operation is performed. So, after performing all these steps, finally we’ve detected the region of interest (ROI), which is the lung tumor.

### 3.7 Feature Extraction

After detecting the tumor successfully, we've extracted several important features of the tumor. They are:

* Area
* Perimeter
* Eccentricity
* Compactness
* And, Circularity

These features are recommended by radiologists and NCCN (National Comprehensive Cancer Network) guidelines. MATLAB determines this features in the ppi or dpi unit. Here ppi stands for pixel per inch, and, the resolution of the CT scanned image is 300 dpi, so we had to convert the ppi unit into millimeter to make it practical.

### 3.8 Simulation Result:



Figure 3. 1: Two of the Lung disease CT images and their Histogram



Figure 3. 2: Cropped image and the median filtered image of the CT images



Figure 3. 3: The Enhanced Images and their Histograms



Figure 3. 4: Edges of the Binary images and the Edge subtracted Binary images



Figure 3. 5: ROI of the Lung Tumor and their Classification

Chapter 4

## Lung tumor classification using machine learning method

4.1 Machine learning:

AI makes AI programming a stride further as it empowers savvy figuring out how to happen inside the segment dependent on past work it did or extrapolations produced using information. The product performs complex dynamic cycles as it comes and gains from past exercises. A concise depiction of the exploration papers dependent on Lung Cancer location utilizing distinctive Machine learning calculations are clarified underneath:

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

Further to get better accuracy we have chosen labeled data from Kaggle [30] 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

### **4.3 Vgg16 (Transfer learning)**

VGG16 is widely recognized as a CNN with good generative capacity. It uses the stack of small convolutional kernels instead of large ones to reduce the amount of parameters considerably compared to Alexnet. It only uses size 3 × 3 as its kernel size [33] [34].

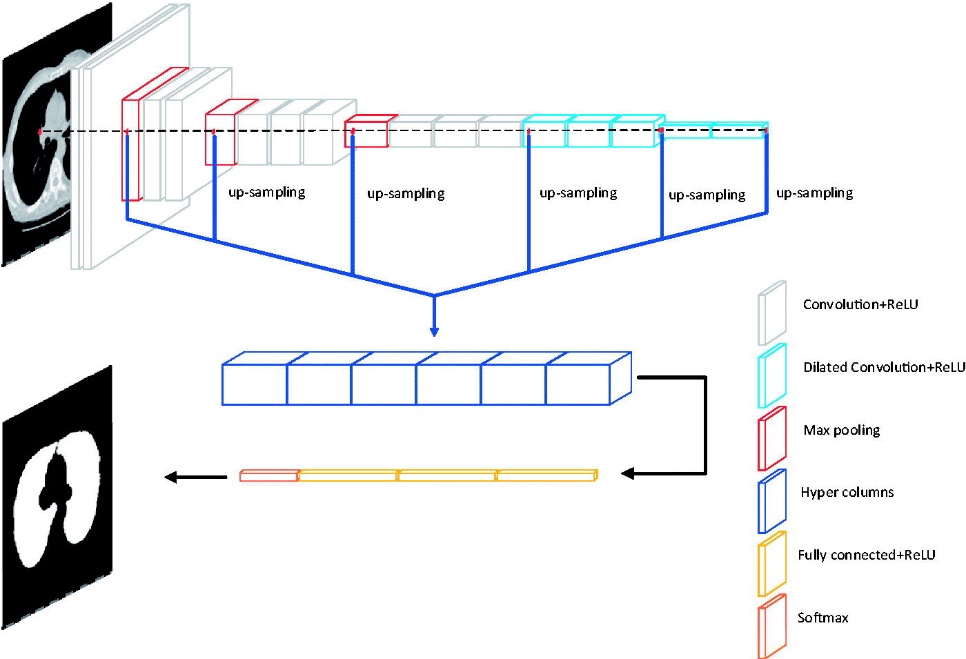


Figure 4. 1: Network Structure

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.

### 4.4 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.

### 4.5 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 [35].



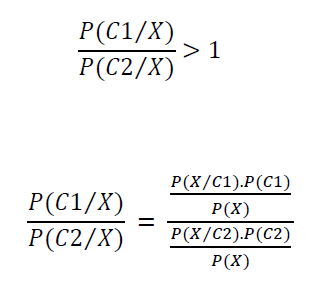
Figure 4. 2: Support Vector Machine

### 4.6 Decision Tree

Decision Trees are a sort of Supervised Machine Learning (that is you clarify what the information is and what the relating yield is in the preparation information) where the information is ceaselessly parted by a specific boundary. The tree can be clarified by two substances, in particular choice hubs and leaves. The leaves are the choices or the ultimate results. What's more, the choice hubs are the place where the information is parted [36].

### 4.7 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 [37]



Figure 4. 3: Example X data and Probability-Class Relations

4.8 Logistic Regression:

Logistic regression is another technique borrowed by machine learning from the field of statistics. the [logistic function](https://en.wikipedia.org/wiki/Logistic_function), also called the sigmoid function was developed by statisticians to describe properties of population growth in ecology, rising quickly and maxing out at the carrying capacity of the environment. It’s an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits.

1 / (1 + e^-value)…..(2)

Where e is the [base of the natural logarithms](https://en.wikipedia.org/wiki/E_(mathematical_constant)) (Euler’s number or the EXP() function in your spreadsheet) and value is the actual numerical value that you want to transform [38].



Figure 4. 4: Logistic Regression to distinguish two classes

### 4.9 Random forest classifier:

Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting [39].

Here in our case the methodology has two phases

1. Preprocessing Phase

2. Prediction Phase

### 1.Preprocessing Phase:

Here we have used two methods in order to preprocess the data.

1. Processed Data

2. Transfer Learning Feature Extraction(vgg-16)

### 2.Prediction Phase:

In this phase, we’re actually using it or putting it use and that is to do a prediction and on there now we have our train model and our new data come together and output is going to be a prediction of what we are looking for.

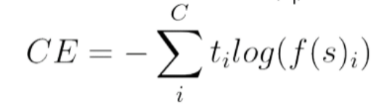
Machine Learning Workflow; It works iteratively;

* Define Objective
* Prepare the Data
* Collect Data
* Select Algorithm
* Train Model
* Test Model
* Predict

### System flow architecture

Figure 4. 5: System Flow 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 machine learning model for the image classification and recognition. Training and testing images were passed through above mentioned algorithms. The softmax function was applied to classify the given object. The model was trained and tested using jupitar notebook .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 softmax 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 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.

### 

### Support Vector Machine on processed Dataset

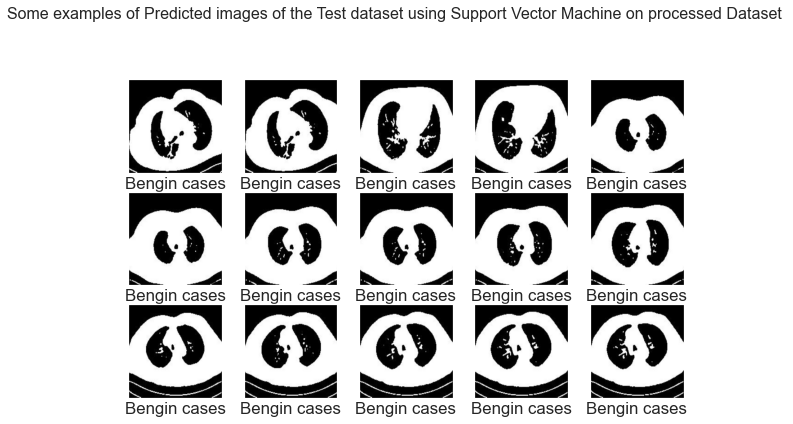


Figure 4. 6: Example of predicted Images of the Test Data using SVM with Processed Dataset

Here the dataset has been used from the test dataset. Support Vector Machine algorithm has detected these images successfully.

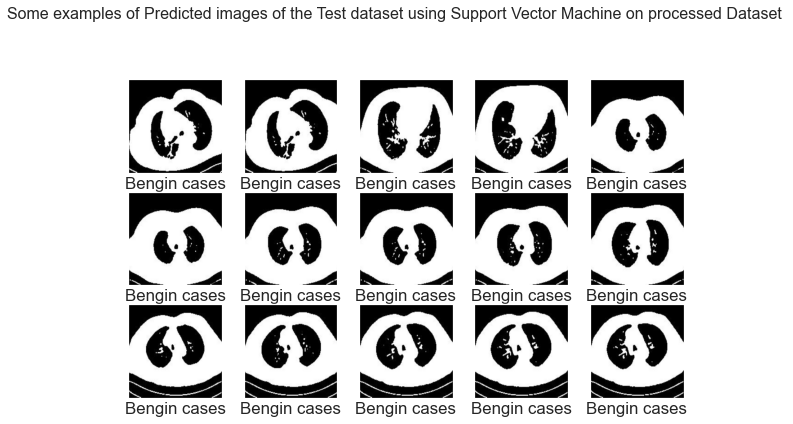
****

Figure 4. 7: Example of Mispredicted Images of the Test Data using CNN without Processed Dataset

Here the dataset has been used from the test dataset. These are the images that the algorithm did not predict correctly.

### Logistic Regression Classifier on processed Dataset

Figure 4. 8: Example of predicted Images of the Test Data using Logistic Regression Classifier on Processed Dataset

Here the dataset has been used from the test dataset Logistic Regression algorithm has detected these images successfully.

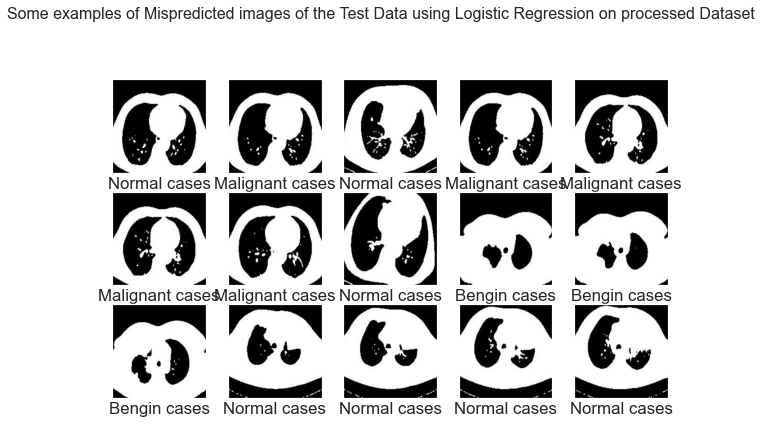
****

Figure 4. 9: Example of Mispredicted Images of the Test Data using Logistic Regression Classifier on Processed Dataset.

Here the dataset has been used from the test dataset. These are the images that the algorithm did not predict correctly.

### 

### Random Forest Classifier on processed Dataset

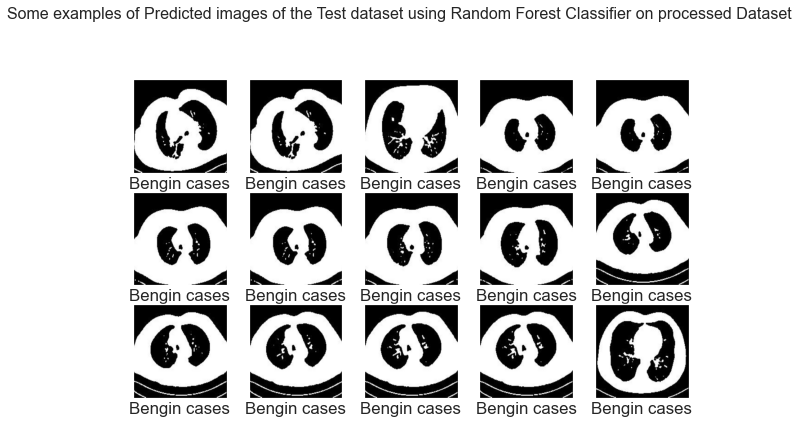


Figure 4. 10: Example of predicted Images of the Test Data using Random Forest Classifier on processed Dataset

Here the dataset has been used from the test dataset Random Forest Classifier algorithm has detected these images successfully.



Figure 4. 11: Example of Mispredicted Images of the Test Data using Random Forest Classifier on processed Dataset

Here the dataset has been used from the test dataset. These are the images that the algorithm did not predict correctly.

### Unprocessed Dataset

### Support Vector Machine on Unprocessed Dataset

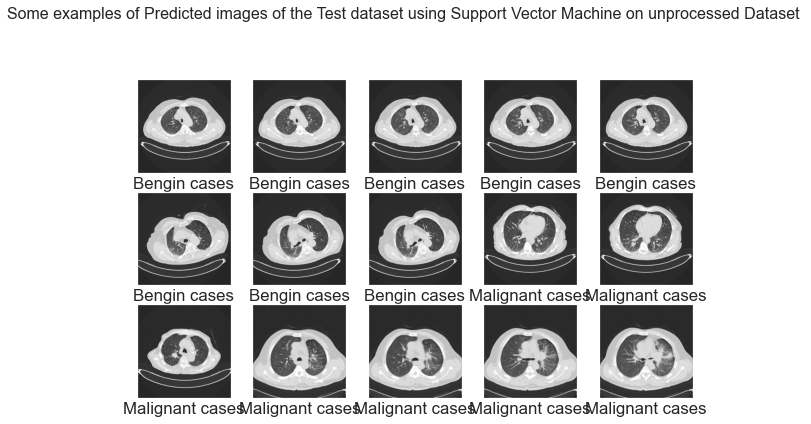


Figure 4. 12: Example of predicted Images of the Test Data using Support Vector Machine on Unprocessed Dataset

Here the dataset has been used from the test dataset Support Vector Machine algorithm has detected these images successfully.

****

Figure 4. 13: Example of Mispredicted Images of the Test Data using Support Vector Machine on Unprocessed Dataset

Here the dataset has been used from the test dataset. These are the images that the algorithm did not predict correctly.

### Logistic Regression on unprocessed Dataset

****

Figure 4. 14: Example of predicted Images of the Test Data using Logistic Regression on unprocessed Dataset.

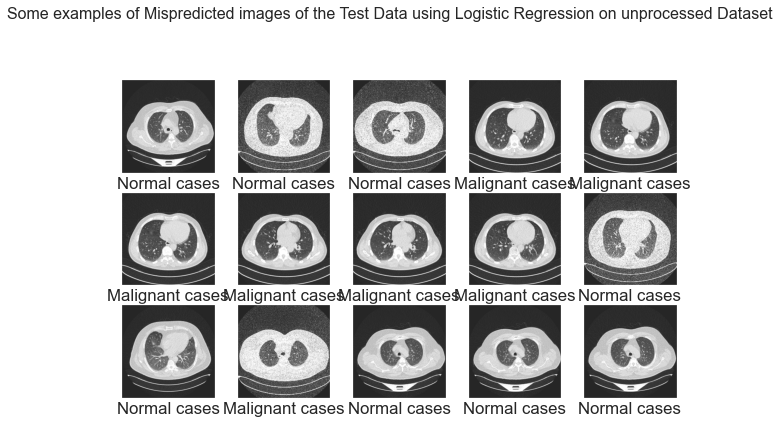
Here the dataset has been used from the test dataset Logistic Regression algorithm has detected these images successfully

Figure 4. 15: Example of Mispredicted Images of the Test Data using Logistic Regression on unprocessed Dataset.

Here the dataset has been used from the test dataset. These are the images that the algorithm did not predict correctly.

### Random Forest Classifier on unprocessed Dataset

****

Figure 4. 16: Example of predicted Images of the Test Data using Random Forest Classifier on unprocessed Dataset

Here the dataset has been used from the test dataset Random Forest Classifier algorithm has detected these images successfully.

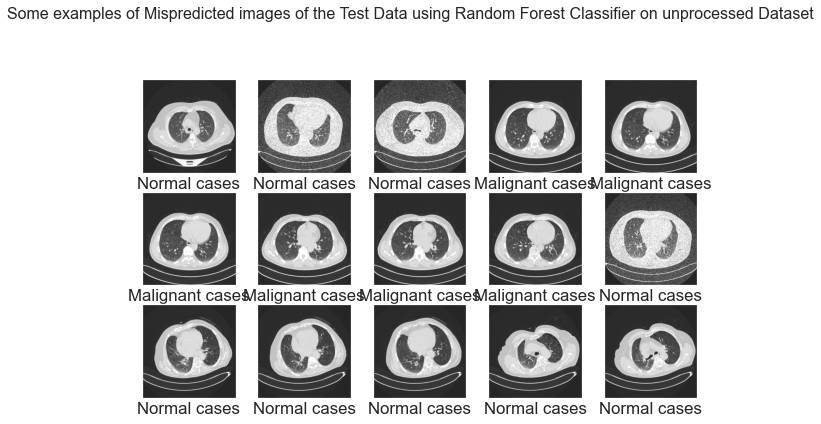


Figure 4. 17: Example of Mispredicted Images of the Test Data using Random Forest Classifier on unprocessed Dataset.

Here the dataset has been used from the test dataset. These are the images that the algorithm did not predict correctly.

### Vgg16 feature extraction with Processed Dataset

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

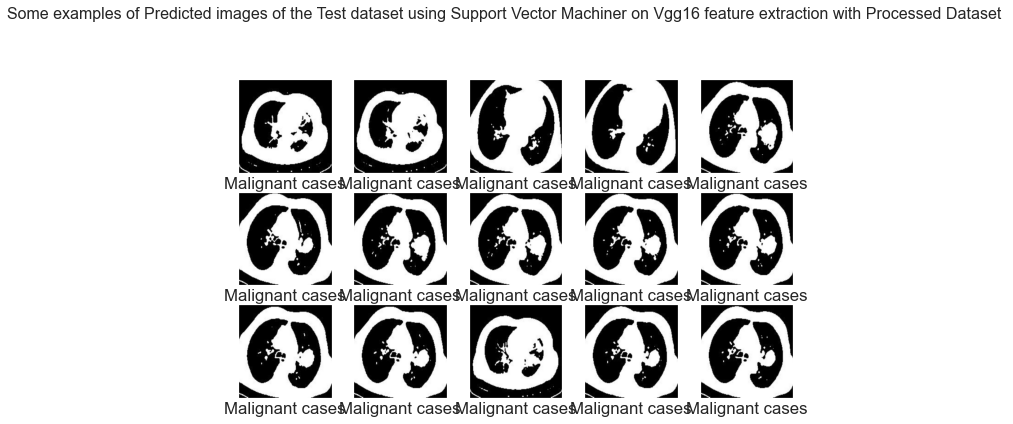
****

Figure 4. 18: Example of predicted Images of the Test Data using Support Vector Machine on Vgg16 feature extraction with Processed Dataset

Here the dataset has been used from the test dataset Support Vector Machine algorithm has detected these images successfully.

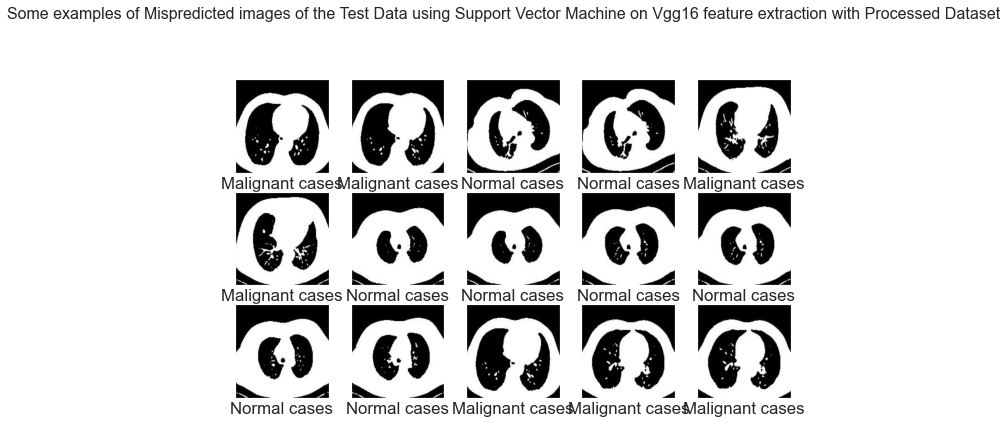
****

Figure 4. 19: Example of mispredicted Images of the Test Data using Support Vector Machine on Vgg16 feature extraction with Processed Dataset

Here the dataset has been used from the test dataset. These are the images that the algorithm did not predict correctly.

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

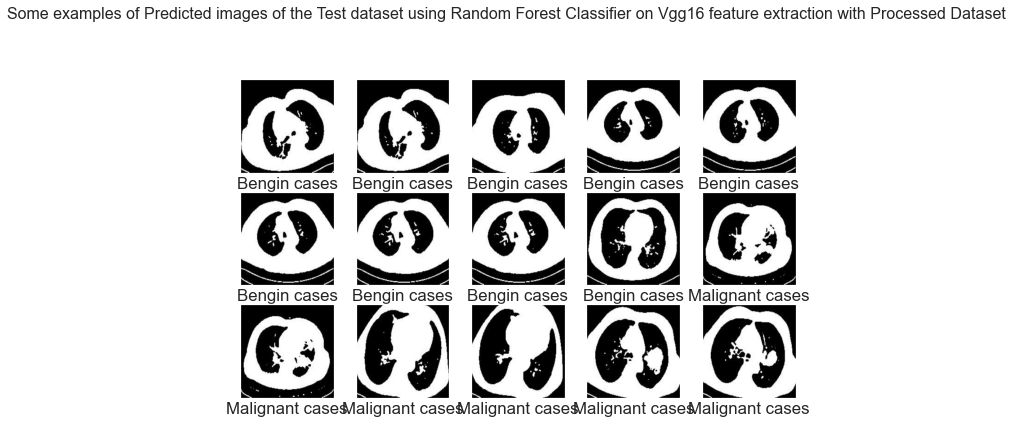
****

Figure 4. 20: Example of predicted Images of the Test Data using Random Forest Classifier on Vgg16 feature extraction with Processed Dataset

Here the dataset has been used from the test dataset Random Forest Classifier algorithm has detected these images successfully.

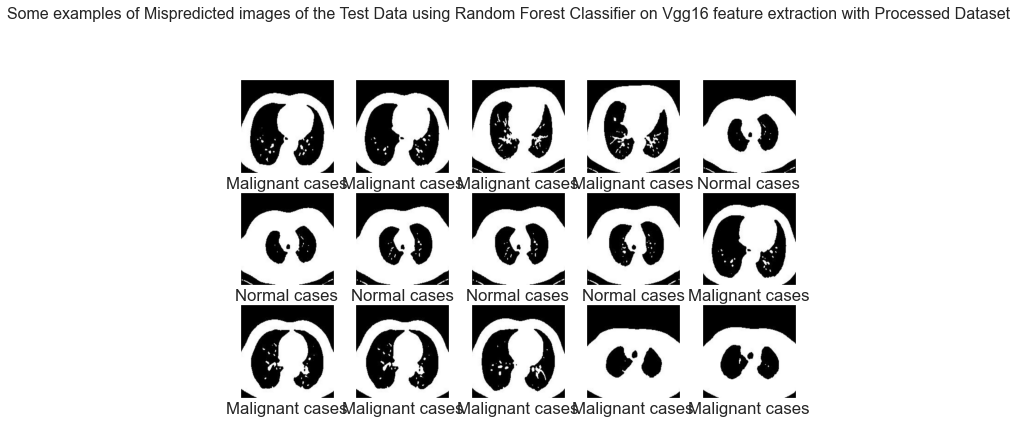
****

Figure 4. 21: Example of mispredicted Images of the Test Data using Random Forest Classifier on Vgg16 feature extraction with Processed Dataset.

Here the dataset has been used from the test dataset. These are the images that the algorithm did not predict correctly.

**Logistic Regression on Vgg16 feature extraction with Processed Dataset**

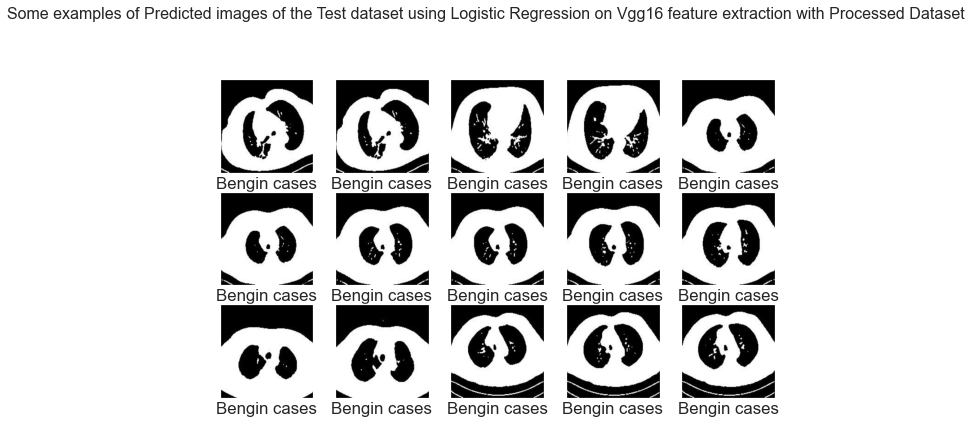
****

Figure 4. 22: Example of predicted Images of the Test Data using Logistic Regression on Vgg16 feature extraction with Processed Dataset.

Here the dataset has been used from the test dataset Logistic Regression algorithm has detected these images successfully

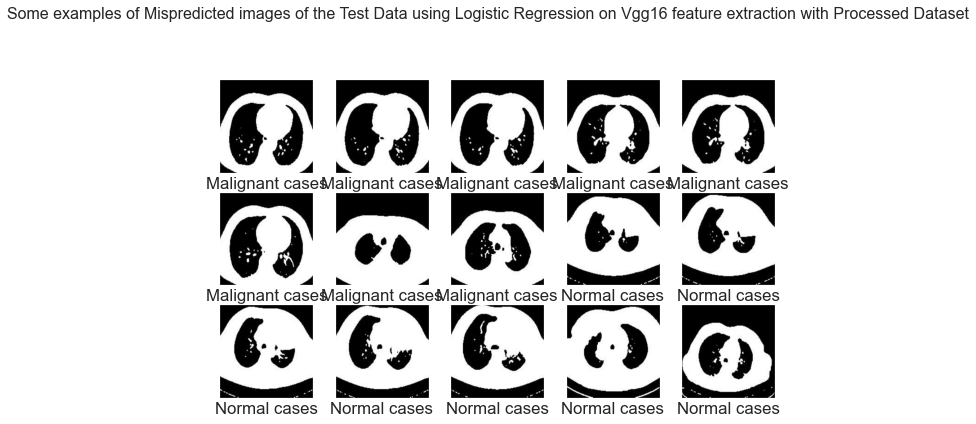
****

Figure 4. 23: Example of Mispredicted Images of the Test Data using Logistic Regression on Vgg16 feature extraction with Processed Dataset.

Here the dataset has been used from the test dataset. These are the images that the algorithm did not predict correctly.

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

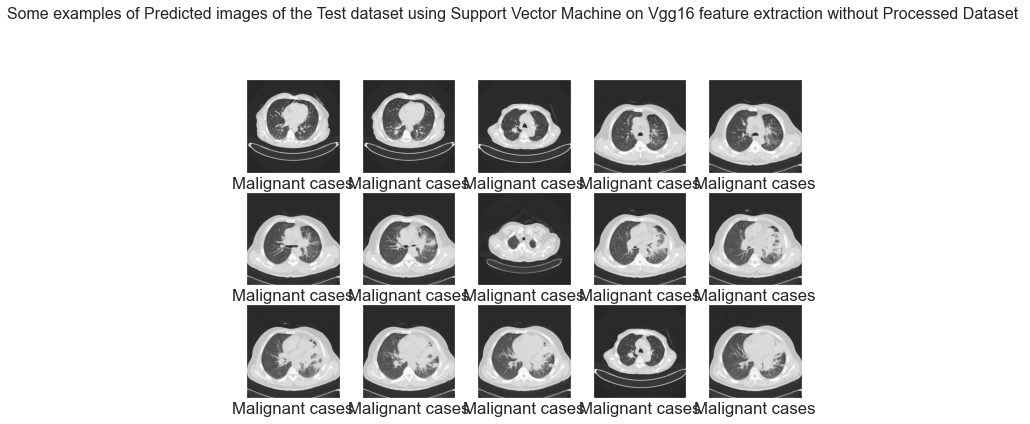
****

Figure 4. 24: Example of predicted Images of the Test Data using Support Vector Machine on Vgg16 feature extraction without Processed Dataset.

Here the dataset has been used from the test dataset Support Vector Machine algorithm has detected these images successfully.

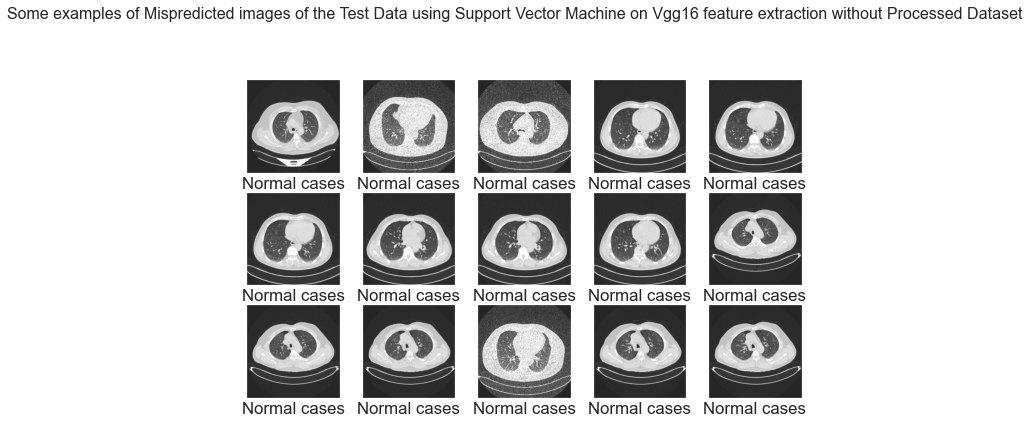
****

Figure 4. 25: Example of Mispredicted Images of the Test Data using Support Vector Machine on Vgg16 feature extraction without Processed Dataset.

Here the dataset has been used from the test dataset. These are the images that the algorithm did not predict correctly.

### Logistic Regression on Vgg16 feature extraction without Processed Dataset

****

Figure 4. 26: Example of predicted Images of the Test Data using Logistic Regression on Vgg16 feature extraction without Processed Dataset

Here the dataset has been used from the test dataset Logistic Regression algorithm has detected these images successfully

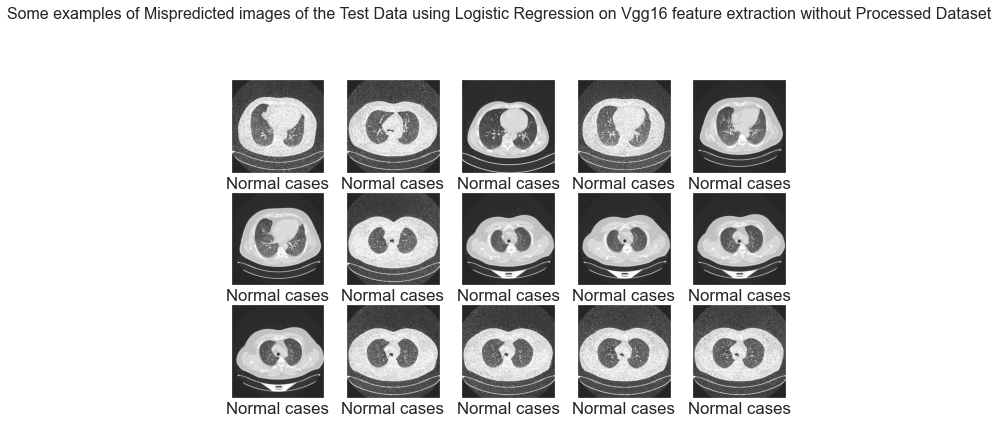
****

Figure 4. 27: Example of Mispredicted Images of the Test Data using Logistic Regression on Vgg16 feature extraction without Processed Dataset

Here the dataset has been used from the test dataset. These are the images that the algorithm did not predict correctly.

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

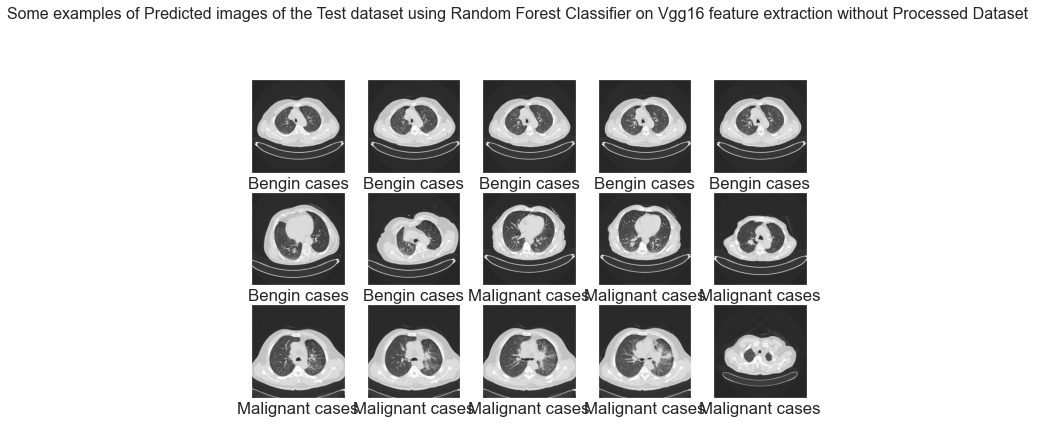
****

Figure 4. 28: Example of predicted Images of the Test Data using Random Forest Classifier on Vgg16 feature extraction without Processed Dataset.

Here the dataset has been used from the test dataset Random Forest Classifier algorithm has detected these images successfully.

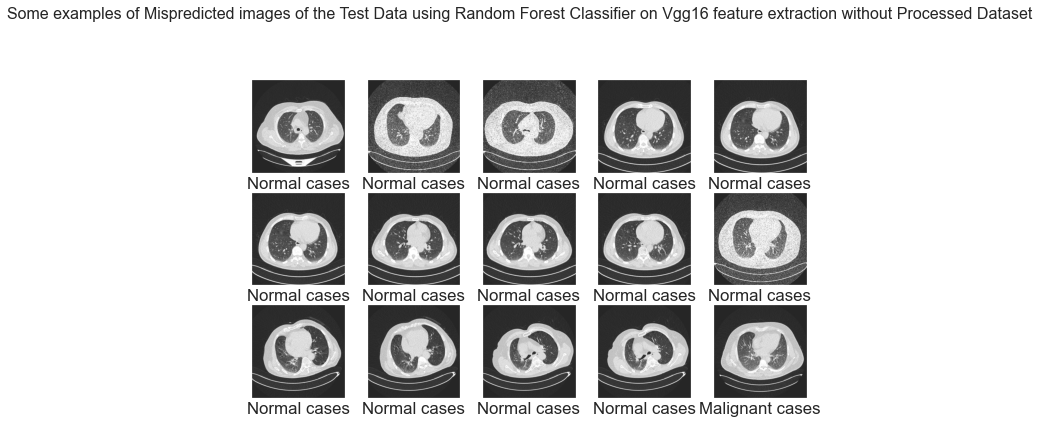
****

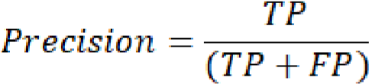
Figure 4. 29: Example of Mispredicted Images of the Test Data using Random Forest Classifier on Vgg16 feature extraction without Processed Dataset

Here the dataset has been used from the test dataset. These are the images that the algorithm did not predict correctly.

### 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.

 …………… (3.1)

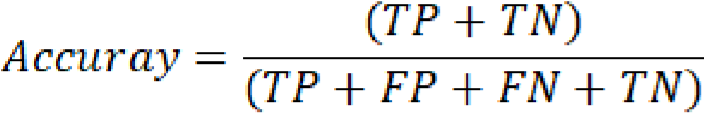
* **Recall:** Recall is the ability of a classifier to find all positive instances. For each class it is defined as the ratio of true positives to the sum of true positives and false negatives.

* **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.



……….. (3.2)

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

 …………. (3.3)

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.

### Benign Classification Result

Table 4. 1: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 |

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. 30: Processed Dataset- benign case classification results

Figure 4. 31: 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. 32: Vgg16 Feature Extraction with Processed Dataset-BENIGN Case Classification Results.

Figure 4. 33: 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**

Table 4. 2: 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 |

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. 34 : Machine Learning with Processed Dataset-Malignant Case Detection Result

Figure 4. 35: 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. 36: Vgg16 Feature Extraction with Processed Dataset-Malignant Case Detection Result

Figure 4. 37: 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**

Table 4. 3: 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 |

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

Figure 4. 38: Processed Dataset-Normal Case Classification Result

Figure 4. 39: Without Processed Dataset-Normal Case Classification Result.

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

Figure 4. 40: Vgg16 Feature Extraction with Processed Dataset-Normal Case Classification Result.

Figure 4. 41: 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 resul

**Chapter 5**

**Lung tumor classification using deep learning method**

### 5.1 Deep learning

Deep learning is a subfield of AI worried about calculations motivated by the construction and capacity of the mind called counterfeit neural organizations.It is a sort of AI and man-made consciousness (AI) that mimics the manner in which people acquire particular kinds of information. Deep learning is a significant component of information science, which incorporates measurements and prescient displaying.

### 5.2 CNN

CNN as an administered profound learning device, CNN is an magnificent decision. This calculation is appropriate for multi-class characterization and parallel arrangement (for instance, foreseeing whether an analytic picture contains a threatening tumor) [31,32].The architecture of CNN architecture is given below.

Figure 5. 1: CNN Architecture

### Input Layer

For input layer we used sequential function. A sequential model is appropriate for a plain stack of layers where each layer has exactly one input tensor and one output tensor.In this case we can start our model by passing an input object to our model so that it knows its input shape from the first.

### Convolutional 2D

The first required conv2D Parameter is the number of filters that the convolutional layer will learn. Depending on the complexity of dataset we select the filter range.we start the range from 32 then 64.we used kernel size (3\*3).

### Relu

The rectified linear function (called relu) has been shown to lead to very high-performance networks.This function takes a single number as an input,return 0 if it is negative.Advantage of this function is it does not activate all the neurons at the same time.it activates only dead neuron.

### Maxpooling

It downsamples the input by taking the maximum value over the window defined by pool size along feature axis.

### Dense Layer

Dense adds the fully connected layer to the neural network. Here we used flatten function which convert two dimension matrix into vector.

### Output

Here we used softmax function we converts a real vector into a vector probability.

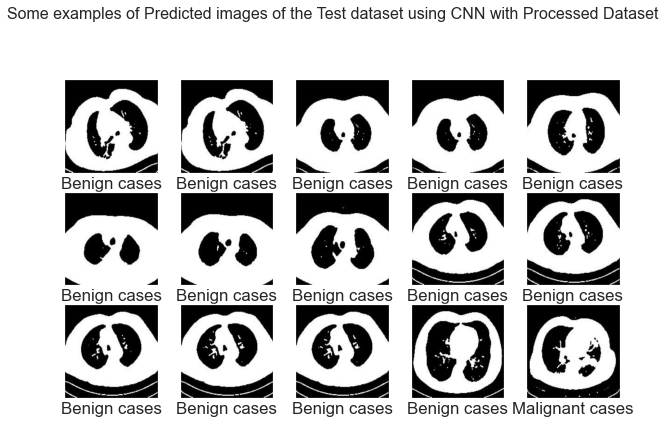
Before passing these stage we first pre-processed our image which we explained in chapter 3, and next we used transfer function for processing image.

To reduce loss we used optimizer which is called adopt optimizer. We also used categorical crossentropy to quantify the difference between two probability distribution.

To monitor the performance of a model during training we used val-accuracy fuction.

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

### CNN with Processed Dataset

**** Figure 5. 2: Example of Processed Image of the Test Dataset Using CNN with Processed Dataset

Here the dataset has been used from the test dataset. CNN algorithm has detected these images successfully.

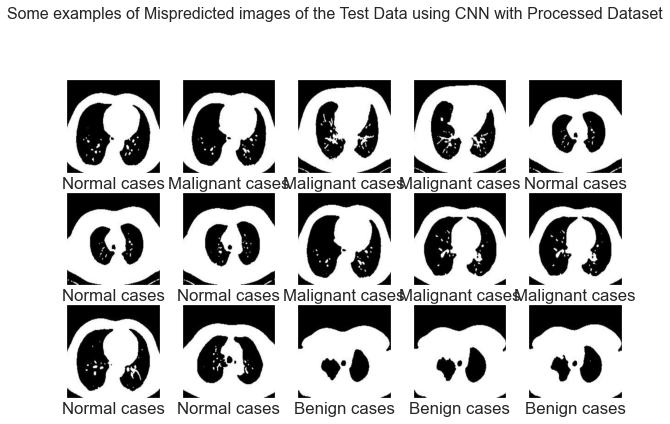
****

Figure 5. 3: Example of mispredicted Images of the Test Data using CNN with Processed Dataset.

Here the dataset has been used from the test dataset. These are the images that the algorithm did not predict correctly.

### CNN without Processed Dataset

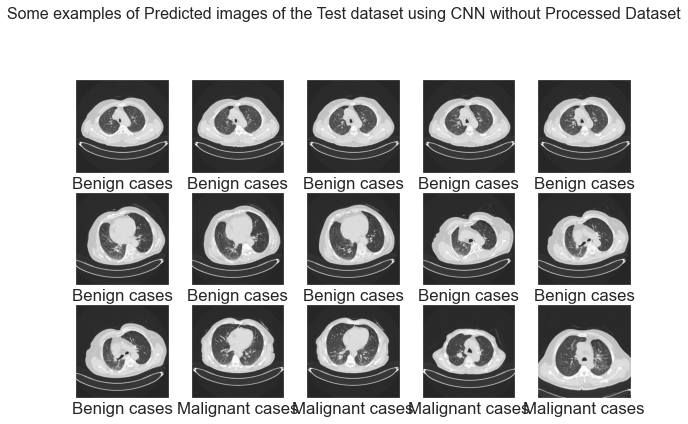


Figure 5. 4: Examples of Predicted Image of the Test Dataset using CNN without Processed Dataset

Here the dataset has been used from the test dataset. CNN algorithm has detected these images successfully.

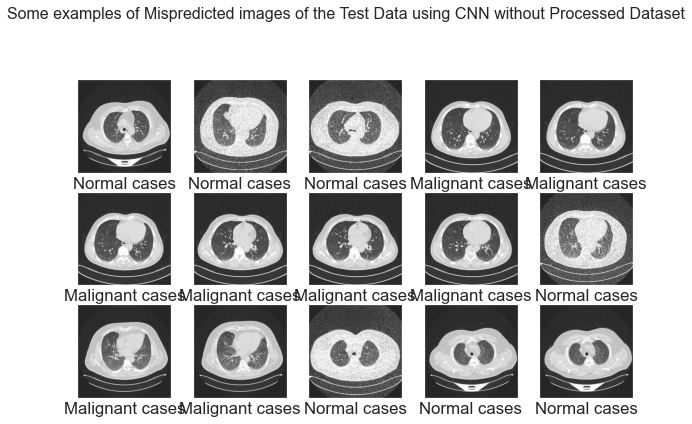
****

Figure 5. 5: Example of Mispredicted Images of the Test Data using CNN without Processed Dataset.

Here the dataset has been used from the test dataset. These are the images that the algorithm did not predict correctly.

### Deep learning (CNN) Classification Results

Table 5. 1: 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 |

### 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 5. 6: CNN with Processed dataset

Figure 5. 7: CNN without Processed Dataset

Figure 5. 8: Vgg16 Feature Extraction with Processed Dataset

Figure 5. 9: Vgg16 Feature Extraction without Processed Dataset.

### Deep Learning Approach

Table 5. 2 **:** 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 |

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. we know as much we will increase our sample validation loss will up and training loss will decrease.This is inverse relationship.

Figure 5. 10: Deep learning (CNN) Approach Accuracy Graph

This is a bar plot of our accuracy Table.

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

Below, Fig. 5.11 and Fig. 5.12 shows the plot of model accuracy vs. epoch and model loss vs. epoch for training and validation images on processed data.

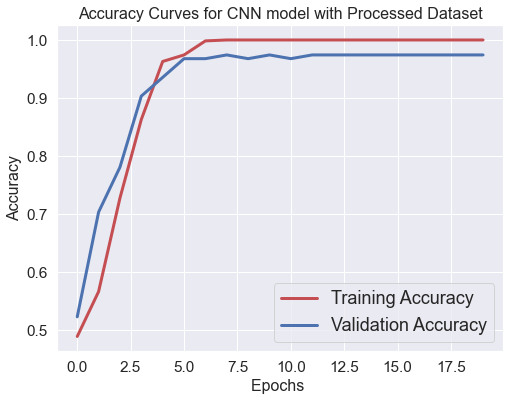
****

Figure 5. 11: Plot of Model Accuracy vs. Epoch for Training and Validation Images.

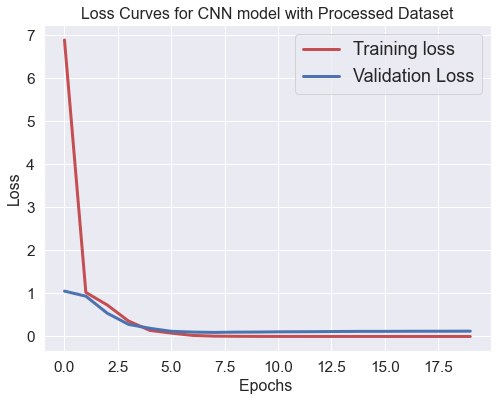
****

Figure 5. 12: Plot of Model Loss vs. Epoch for Training and Validation Images

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

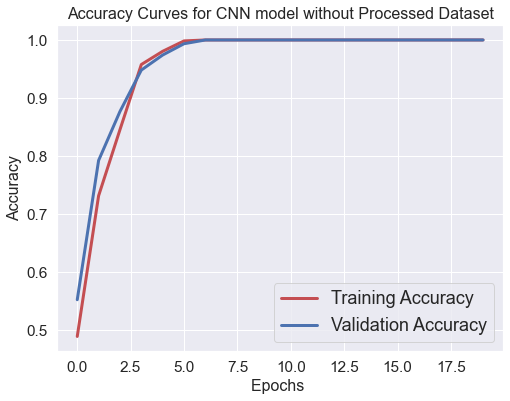
****Below, Fig. 5.13 and Fig. 5.14 shows the plot of model accuracy vs. epoch and model loss vs. epoch for training and validation images on processed data.

Figure 5. 13: Accuracy Curve for CNN model on without Processed Dataset

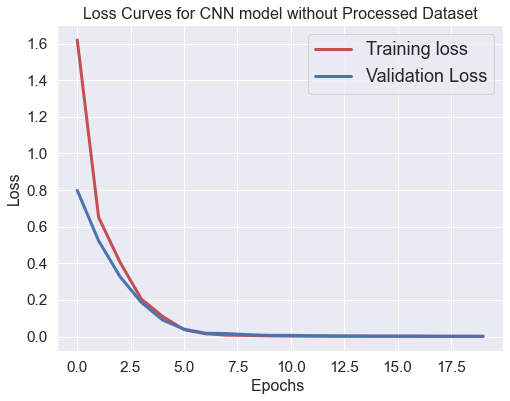
****

Figure 5. 14: Loss for CNN model on without Processed Dataset

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

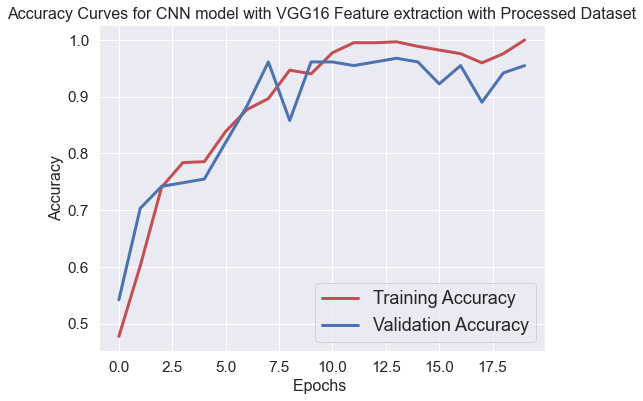
****

Figure 5. 15: Accuracy Curve for VGG16 Feature extraction with processed Dataset

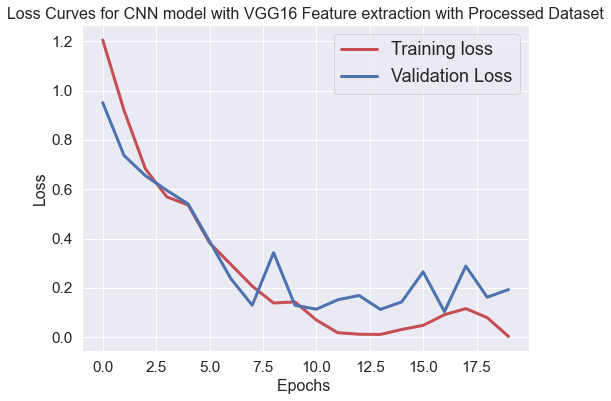
****

Figure 5. 16: Loss Curve for VGG16 Feature extraction with processed Dataset

## 

# Chapter 6

## Monitoring and Analysis of system model

### Accuracy Score comparison with Machine Learning approach

Table 6. 1: 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 |  |

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

Figure 6. 1: Accuracy Score comparison with Machine Learning approach

### Deep learning comparison Table

### 

Table 6. 2: Accuracy Score comparison with Machine Learning approach

### 

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Logistic Regression Classifier** | **Random Forest Classifier** | **Support Vector Machine Classifier** | **CNN** |
| **Without Processed Dataset** | 0.71 | 0.73 | 0.73 | 0.71 |
| **On Processed Dataset** | 0.75 | 0.76 | 0.77 | 0.75 |
| **Vgg16 feature extraction** | 0.82 | 0.83 | 0.88 | 0.84 |

From table 6.1 and 6.2 we can see that with vgg-16 feature extraction from svm algorithms it gives best accuracy, which is 88%. VGG is a more essential engineering which utilizes no lingering blocks. Reset generally perform better then VGG because of it's more layers and remaining methodology. SVM Classifiers offer great precision and perform quicker expectation contrasted with Naïve Bayes calculation. They likewise utilize less memory since they utilize a subset of preparing focuses in the choice stage. SVM functions admirably with an unmistakable edge of detachment and with high dimensional space. The other important advantage of SVM Algorithm is that it is able to handle High dimensional data too and this proves to be a great help taking into account its usage and application in Machine learning field.

## Chapter 7

## Conclusion and Recommendation

### **7.1 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.

### **7.2 Limitations**

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.

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.

### 7.3 Future works

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] | 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. |
| [15] | 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. |
| [16] | T. T. Z Karhan, "Lung Cancer Detection and Classification with Classification Algorithms," *IOSR Journal of Computer Engineering (IOSR-JCE),* vol. 18, no. 6, 2016. |
| [17] | 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. |
| [18] | 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. |
| [19] | 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. |
| [20] | A. A. &. A. Srinivasan, "Deep learning for lung Cancer detection and classification," *springer,* 2020. |
| [21] | 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. |
| [22] | P. BrinthaThereseDr, "Detection of Cancer in Lung with K-NN Classification Using Genetic Algorithm," *Procedia Materials Science,* vol. 10, pp. 433-440, 2015. |
| [23] | 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. |
| [24] | 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. |
| [25] | 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. |
| [26] | 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. |
| [27] | 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. |
| [28] | 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. |
| [29] | 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. |
| [30] | H. KArim, "Kaggle," 24 May 2020. [Online]. Available: https://www.kaggle.com/hamdallak/the-iqothnccd-lung-cancer-dataset. |
| [31] | S. S. Raoof, M. A. Jabbar and S. A. Fathima, "Lung Cancer Prediction using Machine Learning: A Comprehensive Approach," in *2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, 2020. |
| [32] | A. M. A. Jwan Saeed, "Facial Beauty Prediction and Analysis based on Deep Convolutional Neural Network: A Review," in *Journal of Soft Computing and Data Mining, 2(1), 1-12.* , 2021. |
| [33] | Z. G. L. C. Xiaoyun Zhang, "Weld Defect Images Classification with VGG16-Based Neural Network," in *Springer*, 2028. |
| [34] | R. R. S. Dhananjay Theckedath, "Detecting Affect States Using VGG16, ResNet50 and SE-ResNet50 Networks," in *springer*, 2020. |
| [35] | 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. |
| [36] | "xoriant," [Online]. Available: https://www.xoriant.com/blog/product-engineering/decision-trees-machine-learning-algorithm.html. |
| [37] | 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. |
| [38] | "Mechine learning mistry," [Online]. Available: https://machinelearningmastery.com/logistic-regression-for-machine-learning/. |
| [39] | "Javapoint," [Online]. Available: https://www.javatpoint.com/machine-learning-random-forest-algorithm. |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |