

**Subject name: DIGITAL IMAGE PROCESSING (EC662)**

**Event – 4**

**Report on**

**Title: DETECTION OF LUNG CANCER USING DEEP CONVOLUTION NEURAL NETWORKS.**

Submitted by

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sl. No.** | **USN** | **NAME** | **EVENT 2** | **EVENT 4** |
|  | **01JST18EC012** | **ARJUN V SRIVATSA** |  |  |
|  | **01JST18EC088** | **SHREYAS K MAHESH** |  |  |
|  | **01JST18EC102** | **THANMAY M** |  |  |

Submitted to

**SHASHIDHAR R**

**Assistant Professor**

**Department of E & C**

**JSSSTU, SJCE, Mysore**

**DEPARTMENT OF ELECTRONICS AND COMMUNICATION**

**JSS SCIENCE AND TECHONOLOGY UNIVERSITY**

**MYSURU-570006**

**2020-2021**

Acknowledgement

The authors would like to extend their sincere gratitude to their mentor Prof. Shashidhar R, Dept. of Electronics and Communication Engineering, JSSSTU, Mysuru who has rendered his valuable knowledge of this subject and dedicated guidance with a touch of inspiration, vision and motivation. His assistance through giving plenty of early ideas, suggestions and modifications encouraged us to complete the project.

The authors would also thank the HOD of Dept. of Electronics and Communication, Prof. M N Jayaram, the teaching and non – teaching staff of JSS Science and Technology University, Mysuru.

Abstract

Lung cancer is one of the deadliest cancers worldwide. However, the early detection of lung cancer significantly improves survival rate. Cancerous (malignant) and noncancerous (benign) pulmonary nodules are the small growths of cells inside the lung. Detection of malignant lung nodules at an early stage is necessary for the crucial prognosis. Early-stage cancerous lung nodules are very much similar to noncancerous nodules and need a differential diagnosis on the basis of slight morphological changes, locations, and clinical biomarkers. The proposed model consists of three stages which include the visualization of the dataset, applying watershed image segmentation, feature extraction, training the models and classification. The model consists of Luna16 dataset which has around 1000 patient details consisting of all the information, which gets segmented and the image generated is used as the input. Thus, the intention of the proposed work is to classify the cancerous and non-cancerous CT images and to provide with an accurate model.

*Keywords- Lung Cancer, Watershed algorithm, Luna16 dataset, CT images, Deep CNN.*

List of Figures

Figure 1. The block diagram for the proposed system……………………………………..12

Figure 2. The original dicom slices from Luna16 dataset………………………………….13

Figure 3. Visualization of three-dimensional image representation for (a) rib cage (b) lung segments with lung nodules (c) outer lung segment (d) lung nodules…………..14

Figure 4. Metadata information in a single dicom slice……………………………………15

Figure 5. Different markers extracted from CT scan image using watershed algorithm…..16

Figure 6. Image segmentation process visualization……………………………………….17

Figure 7. Datapoints represented on the lung CT image…………………………………...18

Figure 8. Classification accuracy of different models……………………………………...22

Figure 9. The feature extraction graph for different models……………………………….22

Figure 10: The confusion matrix for all the training models………………………………..,23

Figure 11: Deep CNN network architecture…………………………………………………24

Figure 12: U-net Architecture with Batch normalization……………………………………24

Figure 13: Model summary for sequential.1…………………………………………………25

Figure 14: Model summary of sequential\_1…………………………………………………26

Figure 15: Training and Validation accuracy of model sequential…………………………..28

Figure 16: Training and Validation loss of model sequential………………………………..28

Figure 17: Training and Validation Accuracy of model sequential\_1……………………….29

Figure 18: Training and Validation loss of model sequential\_1……………………………..29

Figure 19: The sequential\_1 model training with the epochs. ………………………………30

Figure 20: The sequential model training with the epochs…………………………………..30

Figure 21: The validation set images which has (a) label 1 for cancerous-image and (b) label 0 for non-cancerous………………………………………………………………32

List of Tables

Table 1. Tabular comparison of proposed model with previous models…………………...30

Chapter 1

**Introduction**

This chapter discusses the main concepts that the project is based on and includes overview of the project, objectives, motivation and problem statement.

* 1. **Overview**

Lung cancer is one of the dangerous cancer in the world, with the share of 25% of all cancer related deaths. There are 5 stages in this cancer, and majority of the possible cancer related symptoms will be visible in the last stage. Basically, this type of cancer could be detected with the help of X-Rays, which are low cost and reliable, also CT, MRI, PET scans. But there would be some issues like Privacy of the patient etc. For real scanning using these scans. The concept of Convolution Neural Networks, a technique of deep learning, is the recent concept in the emerging image processing technology. The project involves training CNN model to analyse the obtained dataset, of the patient and provide more accurate result from the dataset. The PET Scan reports as well as X ray radiographs usually comes with the whole lung imaging which is further involved through the pre-processing techniques.

The identification of the diseases in the CT images were employed based on the classification of the images using classification algorithms. Rule-based classifiers were employed which identifies the defects by grouping the pixels that were having similar results for rules used into a group. The rules used in the classifier differs based on the applications for which the process is used. In medical images the statistical parameters and the intensity-based features were also needed to get the best features from the images. The noise that affects the features of CT images, the widespread use of CT image imaging requires the need for developing filter for decreasing noise. Some of the related work is given the most thresholding-based segmentation methods attempt to segment the CT images. Most previous works are prepared to compare different thresholding based image segmentation algorithms based on characteristics such as correctness, stability with respect to parameter choice and stability with respect to image choice. Performance measure like precision, specificity and false positive rate is used to evaluate the accuracy.

* 1. **Motivation**
* In a human body, every process that takes place in the body cannot be predetermined. A stage 0 cancer patient can turn into a stage 1 in a specified period of time, based on the growth of the cancer cells.
* Knowing over the fact that, the behaviour of cancer cells cannot be determined, we acquire a large number of data sets to compensate for the erratic growth of these cells in every scenario.
* This usually leads the model to be trained for a large duration of time, so that the model can assess every scenario presented to it and can correct over its mistakes, since it’s the essence of machine learning itself.
  1. **Problem Statement**

The lung cancer contributes as the second most cause of deaths among various types of cancers nowadays. Early detection of lung cancer cells can be proven as a life saving measure. This involves the application of machine learning and image processing techniques. The common approach deals with the usage of Convolution Neural Network (CNN) in the training of models, with the test and trained datasets. The datasets acquired are usually in the form of X-Ray radiographs, PET scan reports, which is further processed to determine the specific region inside the lungs, where the growth of cancer cells can be observed. The model provides accurate results with precision mapping of the affected regions inside the lungs.

* 1. **Objectives**

1. To help the early detection of different stages of lung cancer by using different advanced technologies of digital image processing.
2. To determine on how the early detection of cancer will significantly reduce the mortality rate by providing appropriate treatment plans for different stages of tumour growth.
3. To analyse the usage of convolution neural networks and image processing techniques which result in parametric optimization for lung cancer recognition from CT images.
4. To effectively train the model using test and train datasets to produce more accuracy in obtaining the results.
5. To make sure that the output data obtained from the fully-trained model will further be reciprocated as valuable resource for the research and development purpose.

6. To make a significant contribution to the world of medical field and to the cancer research studies.

* 1. **Chapter Overview**

The Chapter 1 introduces to the lung cancer, its types and discusses about the problem statement, objective and the motivation behind this project. Chapter 2 talks about the related works in the form of a literature survey that consists of different works related to the proposed model. Chapter 3 discusses about the methodology involved in this project. Chapter 4 plots the results obtained after training the proposed model with the comparison table between proposed work and the existing works. Chapter 5 concludes the project and discuss about the future scope that the project carries. Finally, the references are mentioned.

Chapter 2

**Literature Survey**

This chapter discusses about the related works of the proposed model with its methodology and accuracy.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of the Article** | **Year of Publishing** | **Methodology** | **Accuracy** |
| [1] Lung Nodule Detection based on Faster R-CNN Framework | 2021 | It is used to detect the lung nodules and the training set is used to prove the feasibility of this technique. Has a good accuracy, further assists radiologists in design and development of detection of cancerous cells in lung nodules. | High Accuracy  90% |
| [2] Lung Cancer Detection Using CT Image Based on 3D Convolutional Neural Network | 2020 | In this research, they investigated that 3D CNN to detect early lung cancer using LUNA 16 dataset. The pre-processed raw image is obtained using thresholding technique. Then Vanilla 3D CNN classifier to determine whether the image is cancerous or non-cancerous. The experimental results show that the proposed method can achieve a detection accuracy of about 80% and it is a satisfactory performance compared to the existing technique. | 80% |
| [3] Detection and Characterisation of Lung Tumor by using Convolution Neural Networks | 2020 | Here, The DICOM (Digital Imaging and Communications in Medicine) image of human lungs were taken as the input and various pre-processing techniques were applied for removal of noise, and image segmentation. Finally, the given input images were classified into cancerous and non-cancerous. The possible early detection was also carried out using the Gray Level Co-occurrence Matrix (GLCM). | 88% |
| [4] Mask R-CNN-Based Detection and Segmentation for Pulmonary Nodule 3D Visualisation Diagnosis | 2020 | 3D visualisation diagnosis for pulmonary nodule detection and segmentation is taken up to direct the radiologists for proper diagnosis using M R-CNN technique. Taking the Feature Pyramid Network, for exploring multiscale features and Region Proposal Network (RPN), and finally multiplying the raw medical images to obtain of probable sequences, and ray-casting algorithm was implemented for the above-mentioned 3D models for proper analysis. | 90% |
| [5] Lung cancer detection and classification with 3D convolutional neural network (3D-CNN) | 2017 | A method proposed which involved computer-aided diagnosis (CAD) system for lung cancer classification of CT scans with unmarked nodules, a dataset from the Kaggle Data Science Bowl, 2017. This approach was recorded as the first usage of thresholding segmentation which segmented out lung tissue from the rest of the CT scan. Thresholding produced the next best lung segmentation. | High accuracy and performance rate. |
| [6] Deep learning for lung Cancer detection and classification. Multimedia Tools and Applications. | 2020 | It involves the usage of best feature extraction techniques such as Histogram of oriented Gradients (HoG), wavelet transform-based features. After extracting texture, geometric, volumetric and intensity features, Fuzzy Particle Swarm Optimization (FPSO) algorithm is applied for selecting the best feature. Finally, these features are classified using Deep learning. A novel FPSOCNN reduces computational complexity of CNN. | 87% |
| [7] Early Lung Cancer Detection using Deep Learning Optimization. | 2020 | It involves low dose computed tomography (LDCT) that can reduce the mortality rate by more than 20%, there-for it has been recommended in the typical United State screening. The main advantage of using LDCT screening is its ability to show high resolution chest images with safer x-ray dosage. | High accuracy. |
| [8] Classification of Non-Small Cell Lung Cancer using One-Dimensional Convolutional Neural Network |  | Another journal proposed to develop a one-dimensional Deep Learning Model for automated staging and grading of NSCLC with better accuracy. The objective was also to compare the proposed model with a few leading machine learning methods along with a few notable contemporary investigations. | 73.5% |
| [9] Automated Classification of Lung Cancer Types from Cytological Images Using Deep Convolutional Neural Networks | 2017 | The researchers considered Small cell carcinoma, Squamous cell carcinoma, Adeno carcinoma images in their dataset. The DCNN architecture of 3 convolution and pooling layers and 2 fully connected layers with dropout 0f 0.5 were used. The model developed was able to achieve the overall accuracy of 71.1%, which is quite low. | 71.1% |
| [10] CNN-based Method for Lung Cancer Detection in Whole Slide Histopathology Images | 2019 | Here, histological slides were stained observed and scanned by an automated microscope (with 20x) magnification. Then WSIs were employed to extract the particular region. Some patches created from overlapping were labelled as tumor, if 75% of its pixels was annotated as tumor. Then two CNN architectures were tested for patch classification, either normal or tumor. | 86.4% |
| [11] Multi-path convolutional neural network for lung cancer detection | 2018 | A multi-path CNN architecture was proposed, to automatically detect lung cancer. The model used 3 path, for every receptive field size, modelling distant dependencies, and then were concatenated to lesser the inconsistency among the structure of the nodules. Also, a retraining phase system for tackling difficulties was also deployed. | 91% |
| [12] Classifying Non-Small Cell Lung Cancer Histopathology Types and Transcriptomic Subtypes using Convolutional Neural Networks | 2019 | Deployed a quantitative histopathology analytic framework, in order to identify the gene-expression subtypes of non-small cell lung cancer objectively. Then, a CNN model was applied to identify tumor regions from adjacent benign tissues, and results were obtained. | Achieved areas under the receiver operating characteristic curves (AUC) > 0.935 |

Chapter 3

**Methodology**

This chapter consists of many concepts which include block diagram, procedure, visualization of the dataset, application of watershed algorithm, feature extraction, classification and then the proposed training model is discussed.

**3.1. Block diagram**

The block diagram is divided into five processes which is represented as the figure shown below.

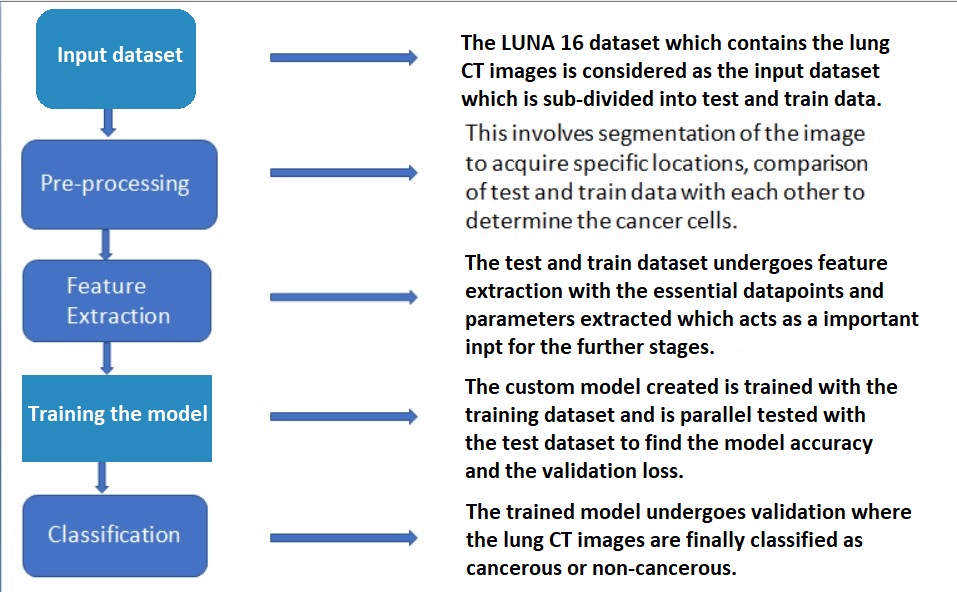
****

Figure 1: The block diagram for the proposed system.

The input dataset in this case, is the LUNA16 dataset which consists of the lung CT images undergoes pre-processing with the noises removed by sobel filters in watershed algorithm. This is further sent to feature extraction where the essential datapoints and the parameters of the dataset images are extracted and stored in a .csv file. The .csv file is taken as an input for training the model with the steps\_per\_epoch, batch size and epochs value set. The obtained trained model with specified accuracy is validated with the CT images to classify which image is cancerous or non-cancerous.

**3.2. Procedure**

**3.2.1 Visualization of the dataset**

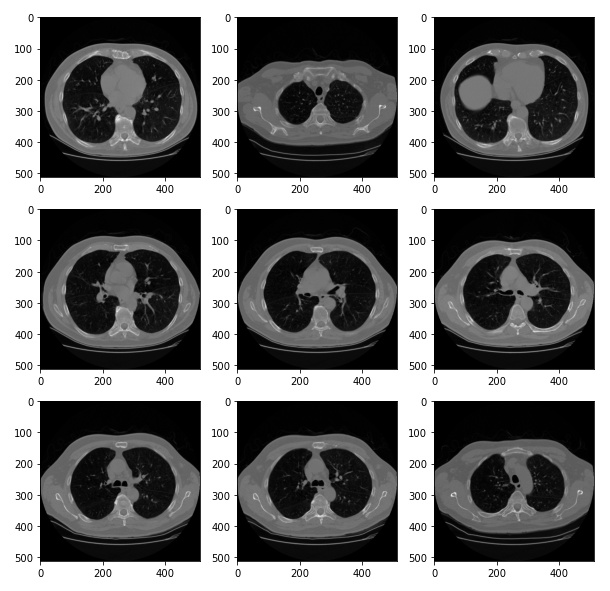
****Visualization of dataset is an important part of training which gives a better understanding of how the dataset is built. But CT scan images are hard to visualize for a normal pc due to its inaccessible format. This problem is solved by importing the pydicom from the dicom image library. The pydicom library gives an image array and metadata information stored in CT images which include the patient name, patient id, patient birth date, image position, image number, doctor name, doctor birth date and many more information which finally is redundant as we only extract the CT image The Luna16 dataset is a directory which contains many subdirectories arranged based on patient id. A complete subdirectory is a three-dimensional image of lungs which is stored in around 180 two-dimensional image slices according to their image number.

Figure 2: The original dicom slices from Luna16 dataset

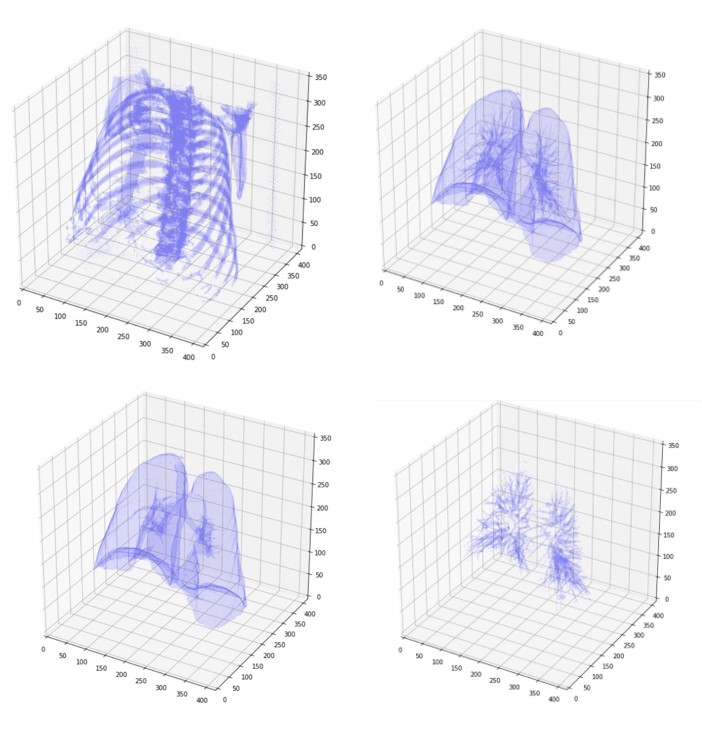
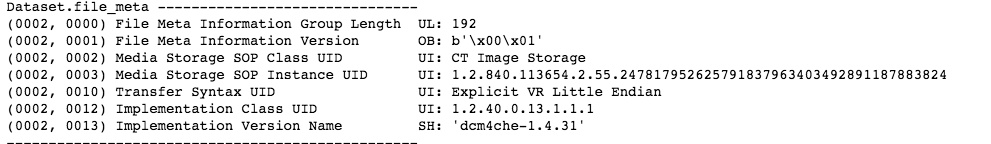
 After combining all the images into a single subdirectory accordingly, it represents a three-dimensional image. The generated 3D image of lungs is essential as it gives a vast information regarding the lung cancer cells and if any, other diseases present in lungs. Figure 3a visualizes the rib cage in which the lungs are present. Fig 3b visualizes the lung segments with the outer lung and the lung nodules present together. Fig 3c visualizes the outer lung segment only. For this particular project, the Fig 3d is taken into account as it visualizes the lung nodules in which the tumor is mostly found.

Figure 3: Visualization of three-dimensional image representation for (a) rib cage (b) lung segments with lung nodules (c) outer lung segment (d) lung nodules.

Metadata stored in CT scan images can be extracted from images with help of pydicom library. Information stored in one image is shown in the figure 4.



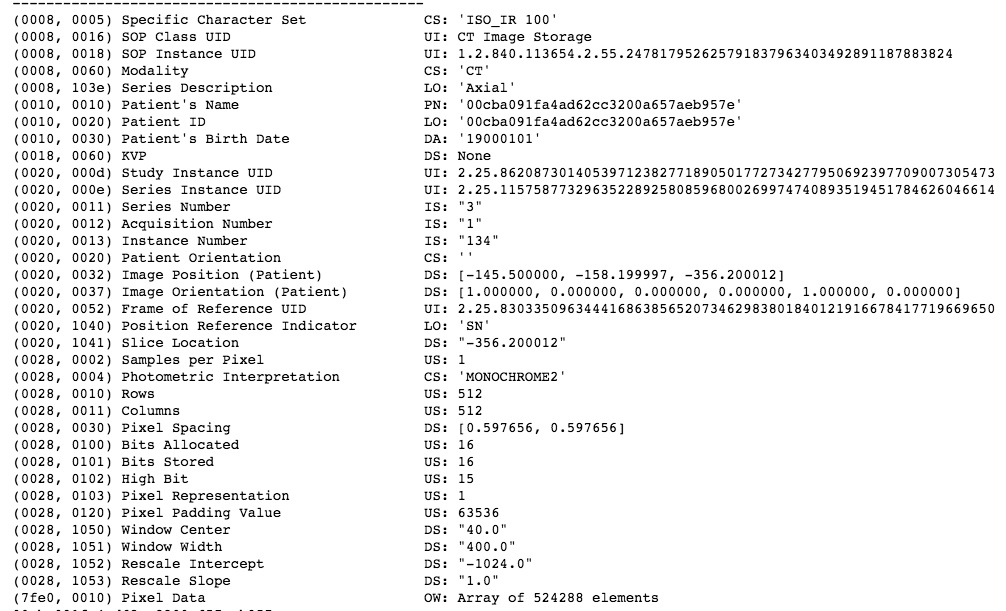


Figure 4: Metadata information in a single dicom slice.

**3.2.2 Watershed Algorithm**

The watershed algorithm is used for image segmentation which is separating different objects in an image. A watershed is a transformation defined on a grayscale image. The name refers metaphorically to a geological watershed, or drainage divide, which separates adjacent drainage basins. The watershed transformation treats the image like a topographic map, with the brightness of each point representing its height, and finds the lines that run along the tops of ridges. The topological watershed was introduced by M. Couprie and G. Bertrand in 1997.

Starting from user-defined markers, the watershed algorithm treats pixels values as a local topography (elevation). The algorithm floods basins from the markers until basins attributed to different markers meet on watershed lines.

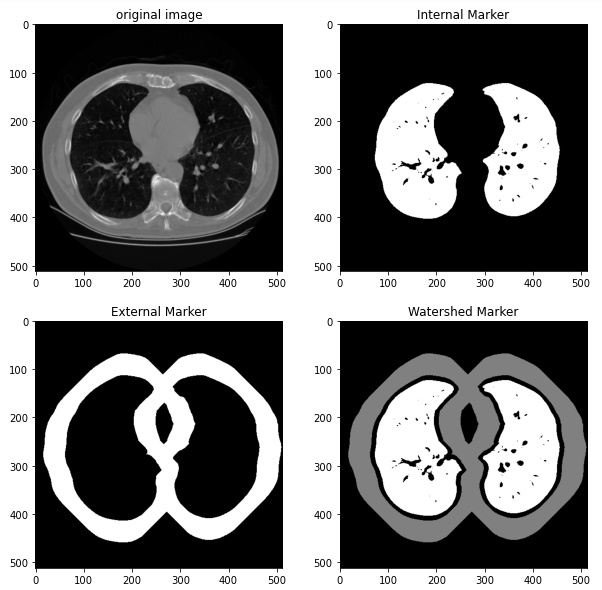
 In this case, the internal and external markers from CT scan images are extracted with the help of binary dilations and added with a complete dark image using watershed methods. This removes external noise from the image and gives a watershed marker of lungs and cancer cells. This is represented in figure 5 where the watershed marker removes external noise and applies a binary mask on the image. Here, the black pixels in lungs represent cancer cells.

Figure 5: Different markers extracted from CT scan image using watershed algorithm.

For better segmentation results, integration of sobel filter with watershed algorithm proves to be useful. It removes the external layer of lungs. After removing the outer layer, the internal marker and the outline that was created to generate the lung filter using bitwise\_or operations of numpy. This also removes the heart from CT scan images. The next step is to close off the lung filter with morphological operations and morphological gradients. It provides better segmented lungs than the previous process. We can see this process in figure 6 below. This process is used to generate a total of 1001 images with related labels which includes almost 12 patients CT scan data in which there is a similar amount of cancer and non-cancer patients.

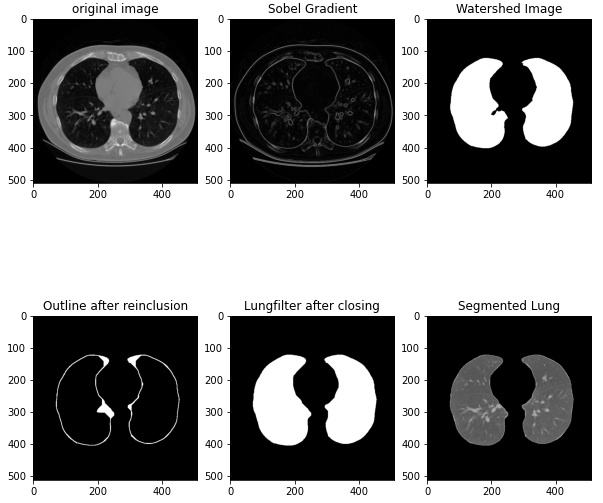


Figure 6: Image segmentation process visualization.

**3.2.3 Feature Extraction**

The images generated by the application of watershed algorithm undergoes feature extraction where the parameters like contrast, homogeneity, entropy, correlation etc., gets extracted and are tabulated in the form of a .csv file. These parameters are obtained by the help of the datapoints which are mapped across the CT image where essential pointers are taken into consideration. These pointers denote the information present in each of them which gets recorded. Figure 7 shows the datapoint extraction in the lung CT image. The generated .csv file is used as an input for the further classification processes.

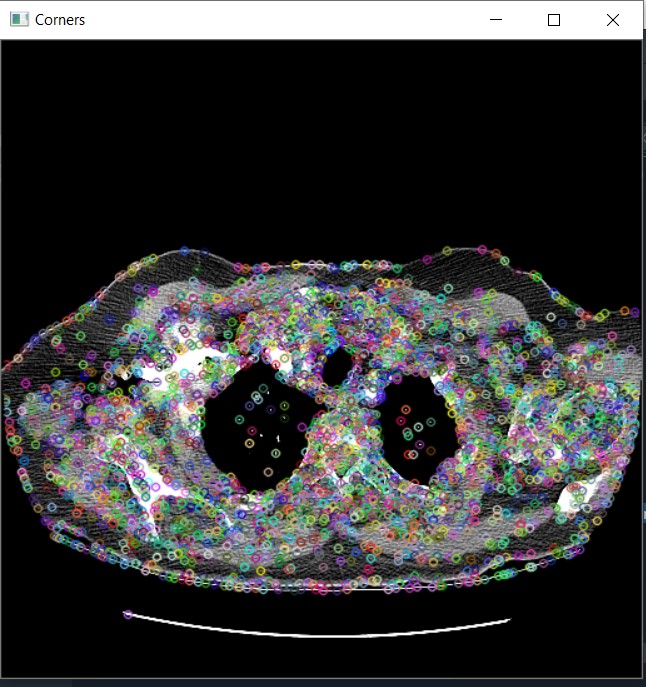


Figure 7: Datapoints represented on the lung CT image.

**3.2.4 Classification based on feature extraction**

The classification part deals with classifying the given input images into either Cancerous or Non-Cancerous. The input fed for the classification part is the dataset of features and characteristics extracted from the Feature Extraction algorithm. This is saved in an Excel-sheet. Here are some codes applied for the starting part of the classification below:

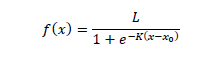
**3.2.4.1 Logistic Regression**

The Logistic Regression Method (LR) is a popular mathematical modelling procedure. It is specially used in analysing the epidemiologic datasets. It is mainly used in the field of machine learning.

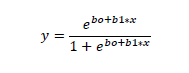
The method can be run in these steps:

1. Calculation using Logistic Function
2. learning of the coefficients for a logistic regression model
3. Estimating predictions using a logistic regression model.

Formula for the logistic function:



The logistic regression model is given in equation as:



E = Euler’s Number

x0 = Middle x - value of sigmoid function

L = The maximum value of curve

K = Abruptness of the curve

(x) = input values and (y) Output value obtained from (x).

**3.2.4.2 Decision Tree Algorithm**

The Decision tree uses supervised technique to build a model which is in the form of a tree data structure. Here the entropy of the parent is calculated. Then, the information gain is calculated by subtracting the weighted sum of entropy of the children from the entropy of the parent. The highest considered Information gain is set as root node, like this the process is done continuously until the classification is done.

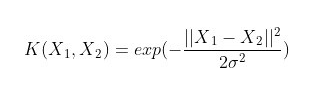
**3.2.4.3 Random Forest Classifier**

The random forest technique is a supervised learning algorithm. The “ forest ” it builds is an ensemble of decision trees, usually trained with the “ Bagging ” method. The general idea of the bagging method is that a combination of learning models increases the overall result. The Random Forest has almost the same hyper-parameters as a decision tree or a bagging classifier. In a random forest, only a random subset of the given features is taken into consideration by the algorithm of a splitting node. The additional usage of random thresholds for each features, can allow making trees more random, rather than searching for the best possible thresholds.

**3.2.4.4 Support Vector Machine (Linear)**

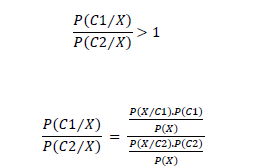
The Support Vector Machine (SVM) is a supervised Machine learning Algorithm, which acn be applied for classification of the given dataset. Here, in this algorithm, each data item is plotted as a point in n-dimensional space (here n = number of features), with the value of each feature being the value of a particular coordinate. Then the classification process is initiated, by finding out the hyper-plane, differentiating the two classes very well.

**3.2.4.5 Support Vector Machine (Radial Basis Function [RBF])**

Typically, the SVM technique used here, can help in classifying linear dataset. Sometimes, the non-linear datasets are also given as an input, but after getting the output, getting the right feature transform will be very difficult to figure out. So in order to solve the problem, the Radial Basis Function kernel is used over the SVM. The formula of a typical kernel is:

**3.2.4.6 Naive Bayes Algorithm**

Naive Bayes Algorithm is used in fields of Data Mining and Machine learning. This is generally used for calculation of probabilities. In order to decide, which class the given instance belongs to, the probabilistic value is calculated. The final class label is the class with highest probability values and vice-versa. This kind of classification using probability is used. Some formulae used



**3.2.4.7 K-Nearest Neighbours Classifier.**

The K-Nearest Neighbours (KNN) Algorithm is a simple and easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems. The KNN algorithm assumes that similar things exist in close proximity, or nearer to each other. It captures the idea of similarity by calculating the distance between the points.

Some Steps to follow here are:

1. Loading of the data
2. Initialization of K value to the chosen number of neighbours, proper selection of the K-value is done to reduce the errors and to maintain the algorithm’s ability to accurately make the predictions.
3. For each instance in the data

* Calculating the distance between the query example and the current example from the data.
* Addition of the distance and the index of the example to an ordered collection

1. Sorting of the ordered collection of the distances and indices from smallest to the largest by the distances.
2. Selecting the first K entries from the sorted collection
3. Getting the labels of the selected K entries.
4. Finally, returning of the mode of the K-labels (Classification Part)

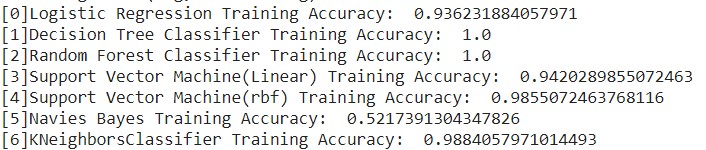
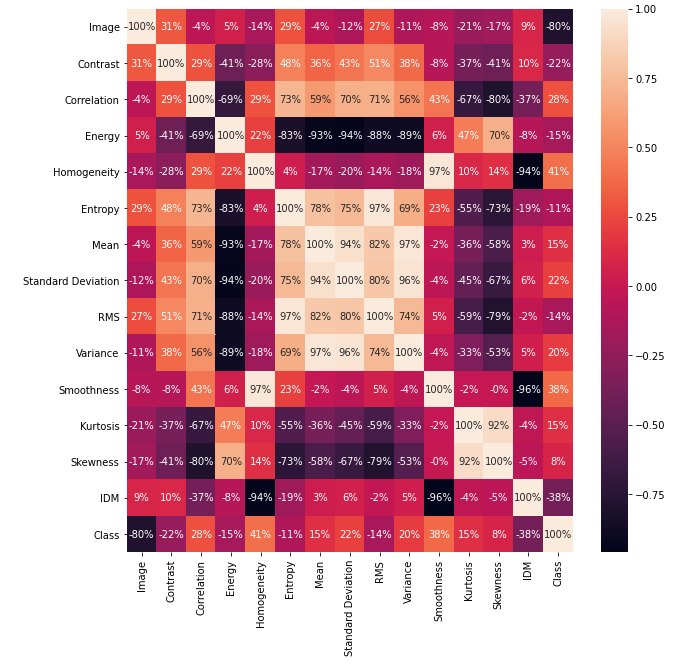
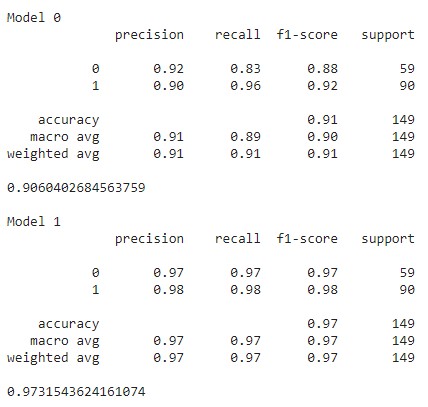
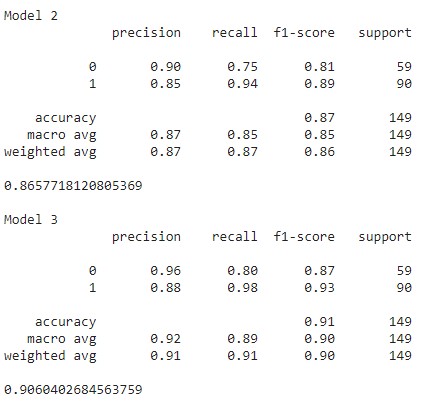
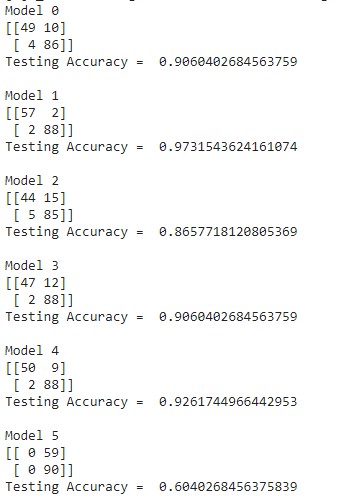
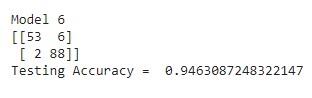
The generated file undergoes classification under 7 models to determine the cancerous and non-cancerous CT images. Figure 8 shows the accuracy of classification in each of the models.

Figure 8: Classification accuracy of different models.

Figure 9: The feature extraction graph for different models.

****

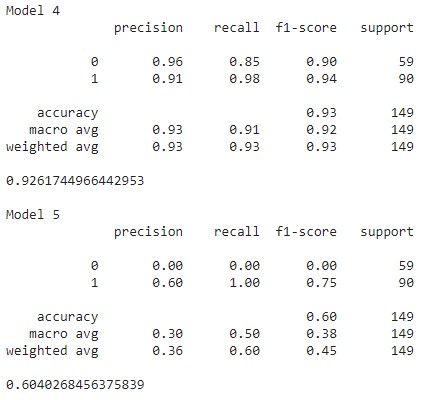
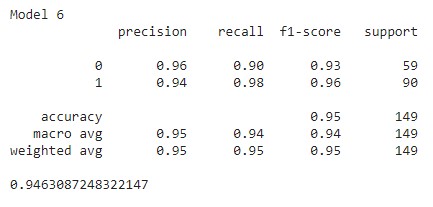
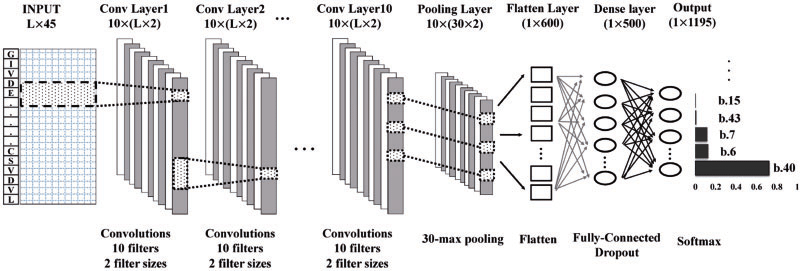
****

Figure 10: The confusion matrix for all the training models.

**3.2.5 Deep Convolutional Neural Network**

In deep learning, a convolutional neural network (CNN, or Conv-Net) is a class of deep neural networks, most commonly applied to analysing visual imagery. They are also known as shift invariant or space invariant artificial neural networks (SIANN), based on their shared-weights architecture and translation invariance characteristics. The study aims to automatically extract the self-learned features using an end-to-end learning CNN and compares the results with the conventional state-of-art and traditional computer-aided diagnosis system’s performance. For the input layer, lung nodule CT images are acquired from the Lung Image Database Consortium public repository having 1018 cases. Images are pre-processed to uniquely segment the nodule region of interest (NROI) in correspondence to four radiologists’ annotations and markings describing the coordinates and ground-truth values. Figure 10 shows the architecture of Deep CNN network.

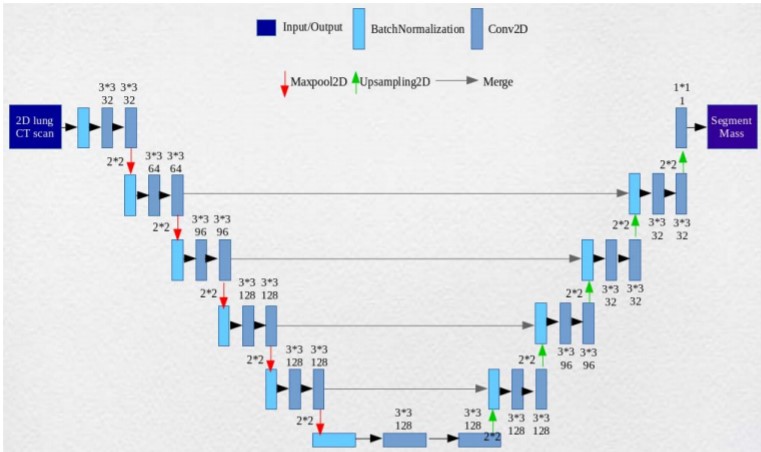
Figure 11: Deep CNN network architecture.

Figure 12: U-net Architecture with Batch normalization.

**3.2.6** **Proposed System Architecture**

The proposed model is a convolutional neural network approach based on lung segmentation on CT scan images. At first, the pre-processing of Luna16 dataset takes place. The next step is to apply lung segmentation with watershed algorithm. Watershed algorithm highlights the lung parts and makes binary masks for lungs using semantic segmentation approach.

The proposed architecture involves two different models of Convolutional Neural Networks, which are based on the comparative study of performance of each type model in different dataset and for different classification problems.

**Case 1:**

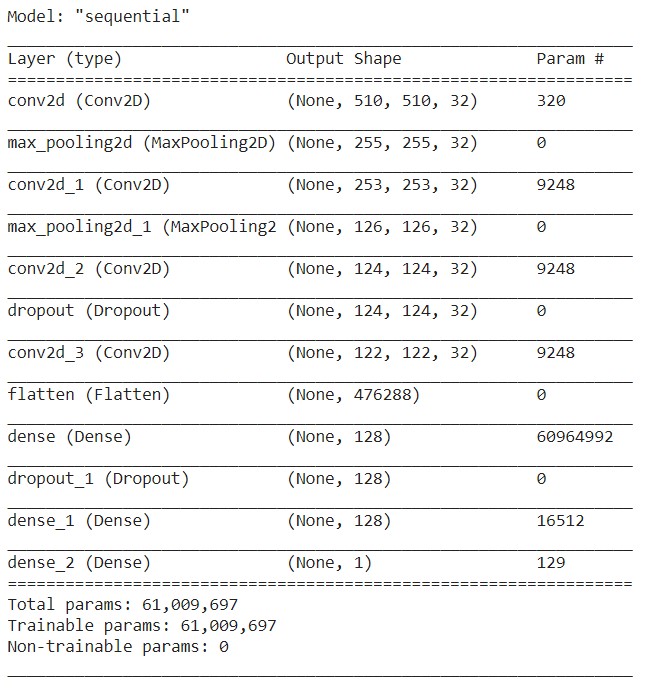
The first model which is named as “sequential” is the basic simple approach of using the convolution layers, flatten fully connected layers, max pooling and dropout in the middle layers, which performs significantly well on the number classification problem. Figure 13 shows the model summary.

Figure 13: Model summary for sequential.

**Case 2:**

The second model is named as “sequential\_1” is the Deep Convolutional Neural Network with max pooling and fully connected layers in the end. This model with specified number of elements and layers performed best in many research papers with different datasets. Figure 14 shows the model summary for sequential\_1.

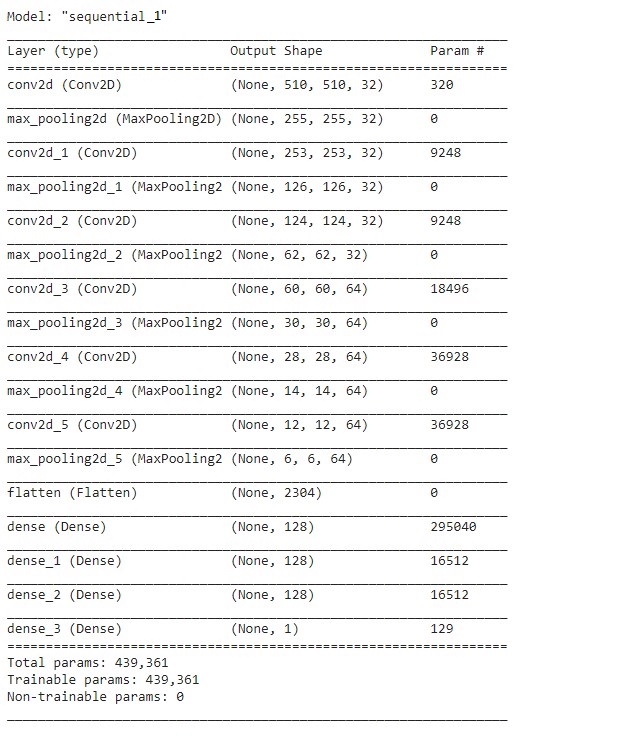


Figure 14: Model summary of sequential\_1.

After making successful binary lung segmented masks, the models are trained on segmented lungs with a batch size of 32 for image data generator. We are training images with the shape of (512,512,1) for the both the models. For a better result, data augmentation is used to train models on different augmentation like shear range, zoom range, horizontal flip, rotation range, centre shift etc. For the end layer, a single node for binary classification is used as to classify between cancerous and non-cancerous lungs. Here, the callbacks from tensor flow keras is applied to save the best accuracy model so that it can run a complete 50 epoch training session to plot the comparison graphs.

Chapter 4

**Results and Discussion**

This chapter include the results obtained from the trained model with the performance comparison between the proposed work and other references.

**4.1 Accuracy and Validation loss graphs**

The following are the graphs of accuracy and loss from the two models used for classification in our project.

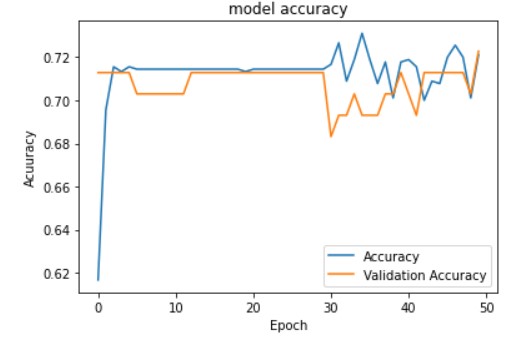
**Model: sequential**

Figure 15: Training and Validation accuracy of model sequential.

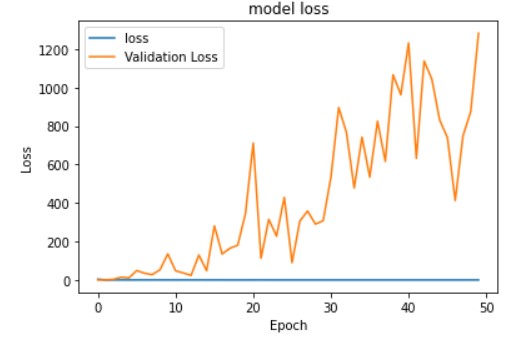


Figure 16: Training and Validation loss of model sequential.

**Model: sequential\_1**

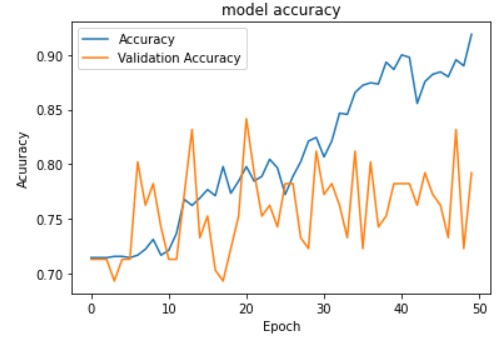


Figure 17: Training and Validation Accuracy of model sequential\_1.

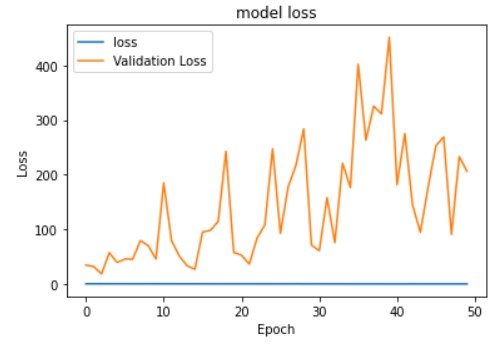


Figure 18: Training and Validation loss of model sequential\_1

The model accuracy and validation loss observed in the sequential model is less than the model accuracy and validation loss observed in sequential\_1 model. The training of the models under 50 epochs are shown in Figure 19 and Figure 20 respectively.

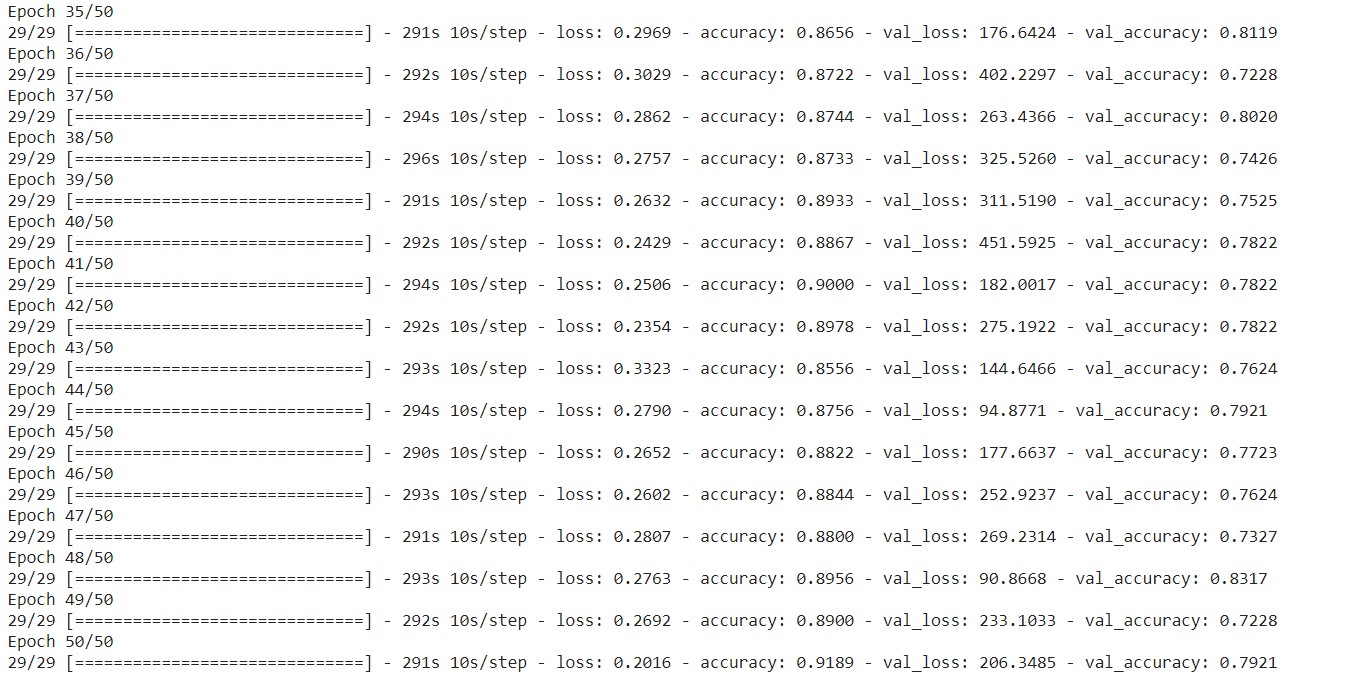


Figure 19: The sequential\_1 model training with the epochs.

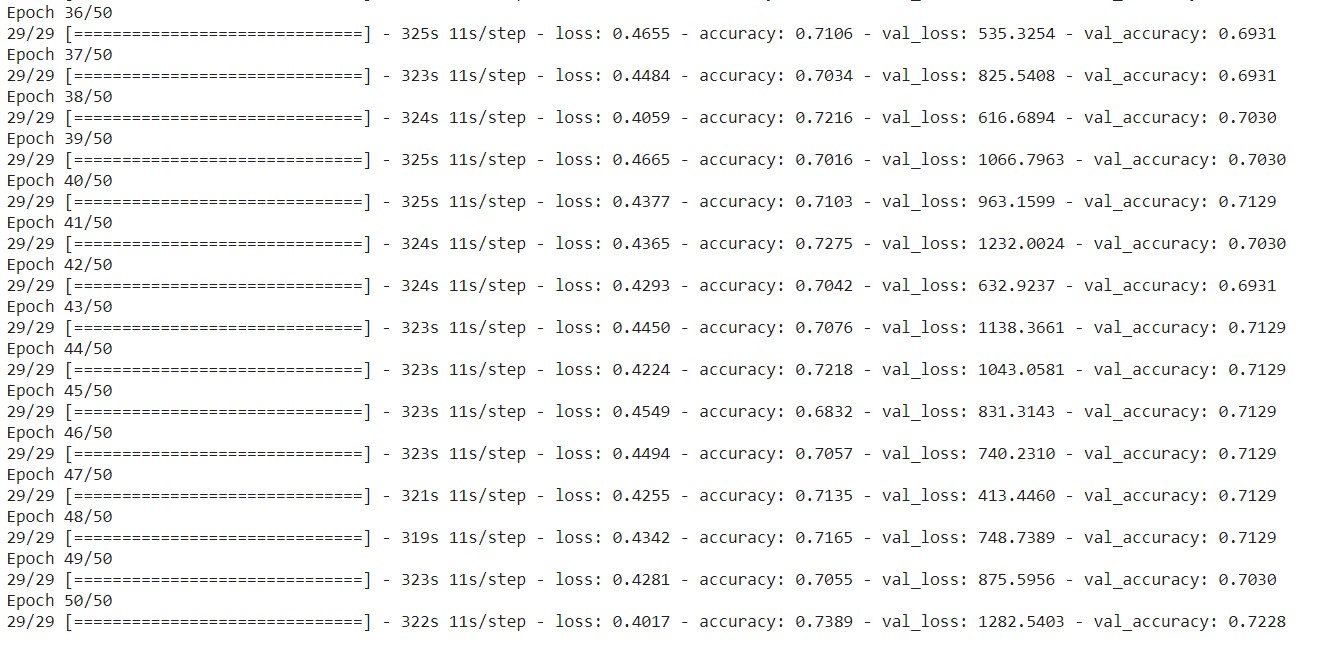


Figure 20: The sequential model training with the epochs.

**4.2 Comparison table**

**4.2.1 Results**

From the results shown in graphs and the comparison table a conclusion can be drawn on the following things:

1. The sequential model is performing the worst overall and even getting over trained without reaching the satisfactory amount of validation accuracy and the validation loss.
2. The sequential\_1 is performing good when training the model and reaching appreciable levels of test accuracy and test loss.
3. The poor performance of the sequential model can be summed up to the fact that the steps\_per\_epoch value has decreased due to the input data running out. This problem is not encountered in terms of sequential\_1 this providing with a better accuracy and validation.

**4.2.2 Performance Comparison**

The table below shows the comparison of the best accuracy model with reference models from other research papers for binary classification. It is noted that the considered 4 models uses the same convolutional neural networks and the rest uses other type of training model. The performance comparison with the same type of training model inferred that the validation accuracy in all the cases is much lower compared to the proposed work. This happens due to reasons which include the dataset difference, the pre-processing algorithms and the training models considered.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Reference works** | **[3]** | **[1]** | **[11]** | **[2]** | **[12]** | **Proposed Work** |
| **Method/Algorithm** | CNN with MATLAB implementation | Fast R-CNN | Multi-path CNN | 3D CNN | AUCs | Deep CNN |
| Dataset | DICOM | LIDC-IDRI | KDSB 2017 | Luna16 | Cancer Genome Atlas | Luna16 |
| Accuracy | 88% | 88.2% | 87.8% | 80% | 93.05% | 92.0% |
| Error rate | - | 20% less than Faster R-CNN | - | - | Better than pathologists data of 87.7% | 26% better accuracy compared to sequential model. |

Table 1. Tabular comparison of proposed model with previous models

**4.2.3 Classification**

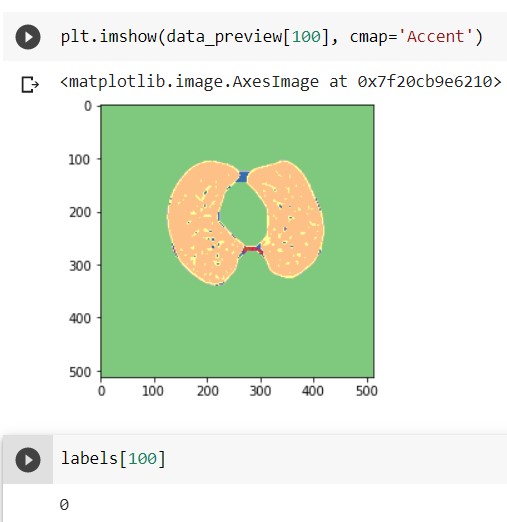
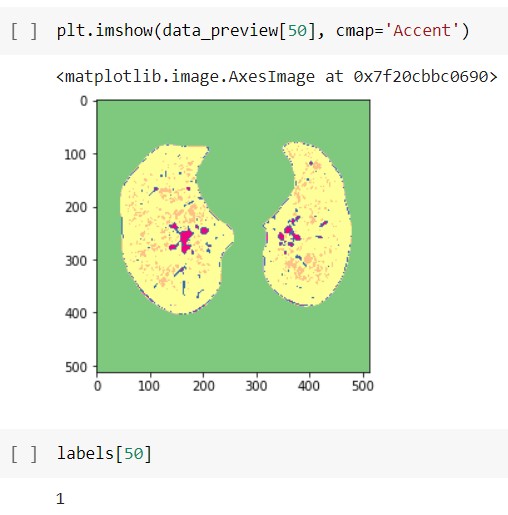
The classification of cancerous and non-cancerous images is done by the labels assigned to it. This is as shown in the figure 21a and 21b.

Figure 21: The validation set images which has (a) label 1 for cancerous-image and (b) label 0 for non-cancerous.

**4.2.4 Challenges**

There were some challenges which was faced during the making of the project which have indirectly impacted the results and outputs obtained. They are:

1. **Dataset:**

The Dataset that we used was in the .dmc format which is different from standard image processing format like .jpg. This format stores the whole information of the patient along with a 3D CT-Scan image of the patient which when converted into 2D numpy image results in the shape of (n, m, m, 1) format which is not supported by the models for transfer learning. This is also in float64 data type which is difficult to handle and process when it comes to CV2, skimage and matplotlib as the standard libraries.

Moreover, when the data is converted into 2D Grayscale image then it becomes difficult to train it with tensorflow and keras framework as they do not support such shapes (n, m, m) during model construction. Now when the images are converted back into RGB or BGR format using different libraries then the output images lose some or many features which are useful in the classification process.

1. **Resources and Hardware Requirements:**

There is a limited number of resources available for the processing of the dataset, first of all the dataset is a hard-to-find resource and after that the hardware for the processing is also limited.

The whole project is made on Google Colab which provides the run time of 12 hour at max for one go and after that it needs a cooldown time of 9-12 hours which limits the progress speed along with the ram storage provided for the processing is 12GB but the when the data is compiled together in a numpy array then it becomes difficult to train small models for more than 1200 images in RGB or even grayscale mode.

Also, when the model is large then due to high number of parameters the terminal crashes as the RAM goes out of memory for regular 32 batch in one epoch for 1000 images so we need to decrease the number of batches which affects the model in more than one way.

Chapter 5

**Conclusion and Future Scope**

This chapter includes the conclusions drawn from the proposed work and the scope of further developments to be made which can increase the validation accuracy.

**5.1 Conclusion**

The deep CNN model considered to be the best approach among the models considered improves the accuracy of the model. The watershed algorithm is applied for the pre-processing of the dataset and image segmentation process. The U-net architecture is used for batch normalization. These components are combined to get the trained model which determines the cancerous and non-cancerous images. The model achieved 93.07% accuracy which is formidable considering the dataset availability and the resource. This work is compared with the existing works and a conclusion was drawn out.

The training period specified for each of the considered model varies as the dataset considered here does not fit into the resource in which the training process takes place. Thus, the first sequential model gets 29 steps\_per\_epoch whereas the second sequential\_1 model gets 100 steps. Both these models are trained for 50 epochs and the accuracy varies as well. There are around 8 other models with the input dataset considered to be feature extracted, to which there are confusion matrices generated as well.

**5.2 Future scope**

* The training model considered can be further developed by adding some more convolutional layers as well as increasing the epochs. This increases the accuracy and the validation loss is reduced considerably.
* Also, many other types of training models can be taken into consideration to make sure that the training gets more efficient.
* The proposed work only determines the cancerous and non-cancerous images which can be further developed to generate the malignancy report and suggest the duration of chemotherapy required for each patient with their unique CT scans.

**References**

[1] Ying Su, Dan Li, Xuedong Chen, “Lung Nodule Detection based on Faster R-CNN Framework”, Computer Methods and Programs in Biomedicine, Volume 200, Science Direct, March 2021.

[2] K. Jalaldeen, M. Malathi, P. Sinthia, M. Vadivel, B.Maruthi Shankar, “An automatic identification of lung tumor by using CNN network and fuzzy-clustering algorithm” Science Direct, 2021.

[3] M. Sangeetha, S. Mythili, “Detection and Characterisation of Lung Tumor by using Convolution Neural Networks”, RASCC 2020, IOP Publishing, 2020.

[4] By Linquin Cai, Tao Long, Yuhan Dai, Yuting Huang, “Mask R-CNN-Based Detection and Segmentation for Pulmonary Nodule 3D Visualisation Diagnosis”, IEEE Access, January 2020.

[5] W Alakwaa, M Nassef, A Badr , “Lung cancer detection and classification with 3D convolutional neural network (3D-CNN) ”, Lung Cancer, 2017 - 13.233.42.234.

[6] Asuntha, A., & Srinivasan, A, “Deep learning for lung Cancer detection and classification. Multimedia Tools and Applications”, 2020.

[7] Elnakib, A., M. Amer, H. & E.Z. Abou-Chadi, F, “Early Lung Cancer Detection using Deep Learning Optimization.”, International Association of Online Engineering, 2020.

[8] Dipanjan Moitraa,\*,Rakesh Kr. Mandalb, “Classification of Non-Small Cell Lung Cancer using One-Dimensional Convolutional Neural Network”, 2019.

[9] T. Atsushi, T. Tetsuya, K. Yuka and F.Hiroshi, “Automated Classification of Lung Cancer Types from Cytological Images Using Deep Convolutional Neural Networks”. BioMed Research International. 2017. 1-6. 10.1155/2017/4067832

[10] Matko Saric, Mladen Russo, Maja Stella and Marjan Sikora, “CNN-based Method for Lung Cancer Detection in Whole Slide Histopathology Images”, 4th International Multidisciplinary Conferences on Computer and Energy Science – IEEE 2019. DOI: [10.23919/SpliTech.2019.8783041](https://doi.org/10.23919/SpliTech.2019.8783041)

[11] Worku Jifara Sori, Jiang Feng, and Shaohui Liu, “Multi-path convolutional neural network for lung cancer detection”, Multidimensional Systems and Signal Processing -SpringerLink 2018.

[12] Kun-Hsing Yu, Feiran Wang, Gerald J. Berry, Christopher Re, Russ B. Altman, Michael Snyder, Isaac S. Kohane, “Classifying Non-Small Cell Lung Cancer Histopathology Types and Transcriptomic Subtypes using Convolutional Neural Networks”, Journal of the American Medical Informatics Association. bioRxiv.org 2019 – DOI: 10.1101/530360.