Lung Cancer Detection And Classification Using Machine Learning

by

Mahbubul Arefin
17201083
Md. Lokman Hekim
18101499
Afia Farjana
19101429
Nisarga Bala
20101533

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
May 2023

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.

2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.

3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.

4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Arefin

Mahbubul Arefin

17201083

Lokman

Md. Lokman Hekim

18101499

Afia

Afia Farjana

19101429

Bala

Nisarga Bala

20101533

Approval

The thesis/project titled "Lung Cancer Detection And Classification Using Machine Learning" submitted by

1. Mahbubul Arefin (17201083)
2. MD. Lokman Hekim (18101499)
3. Afia Farjana (19101429)
4. Nisarga Bala (20101533)

Of Spring, 2023 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on May 25, 2023.

Examining Committee:

Supervisor:
(Member)

Annajiat Alim Rasel

Digitally signed by Annajiat Alim Rasel
DN: cn=Annajiat Alim Rasel, o=Brac University, ou=CSE Department, email=annajiat@bracu.ac.bd, c=BD
Date: 2023.05.22 08:05:23 +06'00'

Annajiat Alim Rasel

Senior Lecturer
Department of Computer Science and Engineering
Brac University

Co-Supervisor:
(Member)

RRahman

Rafeed Rahman

Lecturer
Department of Computer Science and Engineering
Brac University

Program Coordinator:
(Member)

Md. Golam Rabiul Alam,PhD

Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi,PhD

Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

Lung cancer is a term known to all nowadays. This disease grows in the lung tissues and starts to spread with time. The cells responsible for air passage are corrupted by it. It can happen because of air pollution. When we breathe in polluted air regularly, our lungs are likely to be damaged. But by smoking, a lot of people are damaging their lungs repeatedly. Due to this act, they are receiving lung cancer as consequence. It has been affecting people acutely and if prevented in earlier states, then the rate of death would lessen. In order to do that, we have proposed some methods to detect this illness. Machine Learning is a technique where machines (computers) can give us a solution to a problem by analysing the collected data. Using this method, we can detect lung cancer which is the first step towards our desired goal. Usage of CT scan could help us decide between cancer affected and unaffected human cells. Those cells also can be classified more efficiently and we can accurately detect the stage of the cancer when we use CNN models like VGG-19, ResNet50, EfficientNet, DenseNet and so on. We got the highest accuracy from ResNet50 which is 89.52%.

Keywords: Lung Cancer Detection; Machine Learning; Prediction; CNN; CT scan.

Acknowledgement

Firstly, all praise to the Great Allah for whom our thesis have been completed without any major interruption.

Secondly, to our advisor Annajiat Alim Rasel sir and co-advisor Rafeed Rahman sir for his kind support and advice in our work. They helped us whenever we needed help.

Thirdly, to our judging panel faculty Dr. Farig Yousuf Sadeque sir, Md. Tanzim Reza sir, Fairoz Nower Khan ma'am and Zahidul Hasan sir. Though our paper not accepted there, all the reviews they gave helped us a lot in our later works.

And finally to our parents without their throughout support it may not be possible. With their kind support and prayer we are now on the verge of our graduation.

Table of Contents

List of Figures

List of Tables

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

ANN	Artifictial Neural Network
$ASCO$	American Society of Clinical Oncology
CNN	Convolutional Neural Network
$CTScan$	Computed Tomography Scan
DCM	Digital Imaging and Communications in Medicine
LR	Logistic Regression
ML	Machine Learning
MRI	Magnetic Resonance Imaging
PCA	Principle Component Analysis
RNA	Ribonucleic Acid
$SCLC$	Small Cell Lung Cancer
SL	Supervised Learning
SNP	Single Nucleotide Polymorphisms
SVM	Support Vector Machine
VGG	Visual Geometry Group
$WCRF$	World Cancer Research Fund International
$X - ray$	'X' (unknown) Radiation

Chapter 1

Introduction

1.1 Research Motivation

Death is a natural and obvious event in every human life, but if the reason for death is unexpected and caused by some disease, then it is hard to accept. People are dying every year from many diseases and lung cancer is believed to be one of the most frequent ones. This illness is taking a huge number of lives every year. According to ‘World Cancer Research Fund International’ (WCRF), Lung cancer is the 2nd most familiar cancer over the whole planet. More than 2.2 million people will be affected in 2020. ASCO has predicted that more than 2 lakh people will be affected by lung cancer this year in the USA. Almost 69% lung cancer cases are from Asia and 19% deaths are confirmed. The main reason for this happening is not being able to detect this disease at an earlier stage. In the early situation, there are some symptoms in human cells which can help to identify this problem. Not only that, but also it is easier to defeat this disease in this situation as it is comparatively weaker than its next stages. In order to do that, we are proposing some methods by which cancers can be detected in the early stages. Machine learning can be very useful for detecting lung cancer. CT scan is a painless scanning process of the human body which takes minimum 10 minutes to maximum 30 minutes to be completed. Even though it is a lengthy process, the time spent can be quite useful for a patient. It is much better than X-ray results to observe and comprehend the patient’s situation. It can even help us identify the tumors and the blood vessels it feeds from. ‘Chest CT scan’ is a very useful way to detect this illness at an early stage. By CT scanning, very small lesions (cancer affected tissue parts) can be detected. CT scanning is even better than MRI in detecting this disease. It is faster and more accurate. Another proposed method is Support Vector Machine in short SVM, a linear model. It is very useful for regression analysis and classifying data. It can classify between normal and abnormal cells (cancer free and cancer affected cells). With the help of SL and various kinds of algorithms, the classification is done. SL is a branch of Machine Learning where by studying a function, a relation is built between given inputs and outputs. CNN is a type of artificial neural network. It is generally used in image recognition and processing. This method was especially invented to process pixel data. With the help of CNN, we can easily process data and detect whether the cells in the lung are cancer affected or not. Besides with the help of CNN, we can analyze the images and see the differences between the different stages of cancer and successfully detect it.

1.2 Research Problem

Previously many people have died due to misdiagnosis or delayed diagnosis of lung cancer. Due to the symptoms of lung cancer being similar to those of other respiratory disorders including chronic obstructive pulmonary disease (COPD) or pneumonia, it is possible for it to be misdiagnosed or discovered at a later stage. This may lead to delayed treatment beginning and less favorable outcomes. Besides, Lung tumors' ability to be found might depend on their size and location. Imaging scans may overlook small tumors or those positioned in hard-to-reach regions of the lungs, making early detection of them difficult. CT scans and X-rays are the most used diagnostic systems but sometimes even those systems also fail to detect crucial lung tumors. Advances in imaging technology are required to address these problems in a multidisciplinary manner. Thus we are making a custom model based on CNN technology to achieve better accuracy and efficiency. But finding the perfect data set with proper documentation is also tough. Because, most of the datasets available are not pre-processed. Furthermore, we observe that when classifying images of specific classes, even a high-performing model with a very high accuracy might perform poorly.

1.3 Research Contribution

Our research goal has two steps: detecting lung cancer and then classifying the stage. To do that, we ran some existing models with our selected image dataset. The models were CNN models and we chose them as they were suitable for our work. Our research contribution is briefed below for you-

- Understanding the dataset for our research by some steps like classifying images according to their situation (whether the cell is normal, adenocarcinoma, squamous etc.), learning about image height and width in pixels and so on.

- Preparing the dataset for our research by counting the images according to their classes (normal cell class or other), counting the images according to the file types.

- Augmenting the dataset by copping, zooming some images. It was necessary to get better results, otherwise the dataset was becoming biased. Every class had to contain the same number of data, so we made sure of it.

- Then we used our chosen models, one by one on our datasets. It is natural that every model is not going to give us the same result as in accuracy rate. Some gave us comparatively better results, where some results were a bit unexpected.

- Finally, after running the existing models, we tried to propose a model of our own using hyper parameter tuning method. We used this method as it is good in maximizing the performance of machine learning models, helpful for generalizing, improves stability, and enables adaptability to diverse datasets. By fine-tuning there is a possibility to get a very good accuracy if there is enough time.

1.4 Research Objectives

Last but not least, our objectives are-

- Examine current research on the machine learning techniques for detecting and forecasting lung cancer.
- To gather clinical details on individuals with and without lung cancer, as well as imaging data like CT scans.
- To collect the dataset of lung cancer for machine learning and augment the data and preprocess it to get better result.
- To develop and assess several machine learning algorithms' efficacy in predicting lung cancer such as VGG-19, ResNet-50, EffientNetB3, InceptionV3 and so on.
- To evaluate the rended result of the machine learning algorithms in terms of precision, rate of response and accuracy.
- Create a custom made model to get more accurate result and efficiency.
- To identify the most important characteristics and causes of lung cancer development and to determine how they affect the algorithms.
- To assess how machine learning and deep learning could be used to enhance the accuracy of lung cancer detection and forecasting.

Chapter 2

Related Work

Recognizing and identifying lung cancer at the earliest stage is very crucial for any person to survive this disease as it can be incurable once it reaches to stage 2 or stage 3 of cancer. According to the latest WHO published data, it is said that among all different kinds of cancers the most deaths have occurred due to lung cancer which is 35% of all deaths due to cancer. The diagnosis of cancer can be done in various ways but in this modern world of technology, machine learning can be very useful to detect and predict the initial period of lung cancer. There are algorithms of different categories which are used to detect lung cancer, but none of them have 100% accuracy rate. Our work purpose is to find the most explicit algorithm of all and make it even better with higher accuracy rate. In our algorithm, we will detect the cancer affected cell and determine whether it is in the primary, secondary or final stage [4].

Cancer is one of the deadliest and a genetic disorder where the growth of cells in any particular region of our body is unusually high and uncontrolled. There are various types of cancers, but among them, the most affected cancer is lung cancer. According to a recent survey data published by WHO, among all the deaths caused by cancer 40% of it is due to lung cancer. Lung cancer can happen due to various reasons. Deregulation of gene expression alters miRNAs Protein translation steps to generate cancerous proteins. Cigarette smoking causes miRNA deregulation structure. Secondary structure provides the most accurate view of miRNAs. This Structure prediction is very useful to analyze miRNA structural changes. Since the change of miRNAs are easily binding SNPs. These types of comparison can be done more effectively using the machine learning algorithms. The algorithm can easily differentiate between both the structures which can help us in detection of lung cancer and which stage it is in [2].

One of the most severe genetic disorders in the world, cancer, currently has no effective treatment. Among various types of cancers, The highest death rate is 35% for lung cancer. It can happen due to various reasons but it is always better to predict it at the earliest stage to fight against it. We can detect lung cancer using tomography, MRI scans, Chest X-rays, etc. But it is always better to be aware and predict the disease before it spreads. In this advanced era of technology, To determine whether a person has lung cancer or not, we can utilize many sorts of Machine Learning algorithms. Some of the famous algorithms are Naive Bayes, SVM, LR, Artificial Neural Network (ANN), Deep Learning etc. These algorithms and deep learning techniques are implemented to predict lung cancer at the primary stage.

Various types of data such as symptoms based on size and location of a tumor helps us to predict various types of cancer. This research effort presents a concise vision of how we are using a different kind of machine learning algorithms for predicting this illness at its preliminary level[10].

A huge amount of dollars is spent for cancer related medical expenses as it is not detected early. As a result it took many lives every year. Computed Tomography(CT) is very commonly used. Computer Aided Diagnostic(CAD) can help to reduce the pressure of humans, but SCLC is extremely difficult as it looks very similar to one another. A very useful neutral network based algorithm will be Entropy Degradation Method(EDM) which will help to detect cancer cell at a very early stage. The National Cancer institute provided some high resolution CT scans with various sources and qualities which also includes ground level truths, by training and testing data our algorithm's accuracy level is 77.8%. SCLC detection will give a binary output where 0 means healthy patients and 1 means lung cancer developed or developing lung patients, in other words, unhealthy patients. EDM has a good prediction accuracy since it gives a histogram feature for detecting SCLC as well as it has a huge scope for improvement[3].

Lung cancer has a relatively high mortality rate (about 1.3 million people per year), making it the leading type of cancer that kills men and the second type of cancer that takes female lives(as there are comparatively more male smokers). This is because lung cancer cannot be discovered early and can occasionally have no symptoms at all. Early detection can save many lives. There are many algorithms but SVM, ANN, BPN, GA, LDA are more preferable as SVM has 97% accuracy rate, ANN has 96%, DT has 77.5% and so on. For data and processing, Standard Digital Image Database is used where 247 CT scans are used. Principle component analysis(PCA) is used for finding new patterns in high dimensional data as well as it is unsupervised linear conversion technique. K-Nearest Neighbors is very effective, a simple classification method for measuring distance. Support vector machine(SVM) is statistical learning theory and a supervised machine learning algorithm with high accuracy and it works for regression problems. Naive Bayes is a probabilistic implication and It has a better rate of accuracy in order to other classifiers. A decision tree is used when data are being divided simultaneous by any parameter. Artificial Neural Network(ANN) is usually used for image classification problems. Here, PCA, KNN, and SVM are used. In future, different types of methods such as noise cancellations will be used for higher accuracy[6].

The diagnosis of lung cancer brought on by malignant lung tissue is exceedingly challenging since it does not show any symptoms from the early days to present day. By diagnosing it early, lots of lives can be saved. To solve this Neural Network model will be very useful. EDM and ANN algorithms can detect SCLC very early and efficiently. The Neutral Network Back-propagation method has the accuracy over 80%. By combining watershed segmentation and SVM, we get a more accurate picture. For this proposed system approach there are 6 phases to get the output more accurately. Such as- image processing, image filtering, feature extraction, segmentation, edge detection, feature recognition. For proceeding CNN is used with 4 layers (convolutional layer, pooling, flattening, fully connected layer). By doing various relevant work, it has been discovered that a profound neural system approach will produce results that are more accurate[9].

Lung Cancer causes for the uncontrolled growth of a cell in lungs. The main object

of our paper is an early detection of lung cancer by using classification algorithms like as Naive Bayes, SVM, Decision tree and Logistic Regression. Having regular tobacco, excessive use of radon gas, air pollution are major reasons to cause lung cancer. First of all, we have designed the whole classification model like collected dataset from UCI and the data world then we have used those classification modes mentioned before. After performing test we conclude the final accuracy rate of the model. SVM is a supervised learning method which is used for founding the lowest distance sides from the class and trying to find a strengthen space. Then we use a decision tree model which operates supervised learning technique. Naive Bayes mainly used for analyzing large datasets, we have used for four equations regarding Logistic Regression (LR) which is a statistical model which deals with the endemic datasets. After performing the whole operation we have found the result that the highest accuracy rate among all those classification models is SVM which is more than 95%[11].
Nowadays, among the extremely dangerous causes of death all over the world, Lung Cancer is one of the few. According to WHO, if we want to save life, we have to address it as soon as possible. In this paper we have shown, how we get the best possible results using Vector Machine(SVM), K-Nearest Neighbor(KNN) and, Convolutional Neural Network (CNN). After doing so many researches, Researchers have found 'The Using of Machine Learning' for medical diagnostic can be more effective to detect lung cancer earlier. We can categorize lung cancer into two groups- one is small cell lung cancer and another one is the non-small cell. Small cell lung cancer spreads more quickly, which is related to smoking. Machine learning is a child field of AI which has three categories. Supervised Learning is a more effective way as it works in a specific range. We have used materials like dataset which accommodate data regarding clinical forms of lung cancer. We also used a classification model that mentioned before SVM, KNN and CNN. Our main goal to use SVM is to survey data and detect the pattern of lung cancer. KNN is used for find out the difference between research by product and a database element. Lastly, the algorithm of CNN is used for multi-class classification and binary classification which solves the issues like detecting large range of patterns and picture recognition. Finally, we have created a performance evaluation matrix which is a square matrix. The output of detection class is shown in the columns on the other hand, pathological results representing in the rows. After all the research we have found the best results from SVM which is 95.56% and from CNN and KNN we found 92.11% and 88.40% respectively[8].
One of the most suffering and painful deaths by diseases is lung cancer over the whole world and this disease is killing people in a massive rate. The main reason for this many deaths is lack of awareness as it is not detected in time, thus unknowingly the human cells get increased uncontrollably then become weaker as well. So, to stop this disease, we decided to take a step to perceive this illness and we propose a method called 'SVM' and also CT image analyzing. SVM is known as Supervised Learning Method which can classify between normal cells and abnormal cells (cancer affected cells). For CT images we shall check the images and after evaluating the images, blood tests and other reports we can understand the current conditions of the patients. The reports will help us to realize about how much critical situation they are in. Our methods can give us an accurate test results which will help us to move forward with better judgement[5].

Lung cancer is the disorderly abnormal growth of tissues which occurs in the body specifically in airway cells such as lung. Generally, Lung cancer can be divided among two groups by their cells and 4 types of cancer. Pleura Carcinoma is the most killing cancer around the world as it is not detected at the early stage. There are different types of methods is used to identify the cancer to improve the accuracy rate. The accuracy rate by using different types of classification method increased to 98.30%[12].

Chapter 3

Methodology

First of all, our purpose of doing this thesis is to detect lung cancer (if already exists) in the human body. Secondly, there are stages of this cancer, for example: first, second and third stages. We also want to classify the cancer situation by identifying the current stage. For that, we used data set of images of cancer affected lungs. After that, we tested the accuracy of cancer detection of various models by running them with the data set. We used ResNet50, EfficientNet-B7, EfficientNet-B3, VGG19, Inception V3, and MobileNet Version-2 models to test the data set. We also had to augment the data set to increase the amount of images to get a better accuracy rate. For that, we had to zoom, crop, split some images.

3.1 Data Collection

The dataset was collected from kaggle, uploaded by Mohamed Hany. It was a chest CT image dataset with various classes of images. It was a bit prepared to use for CNN. The images were not in dcm format, but in jpg and png format. This data set contains three different subsets: test, train and valid, which represent testing, training and validation sets respectively. Each subset has four types of situations: adenocarcinoma, large cell carcinoma, squamous cell carcinoma and lastly, normal cell.

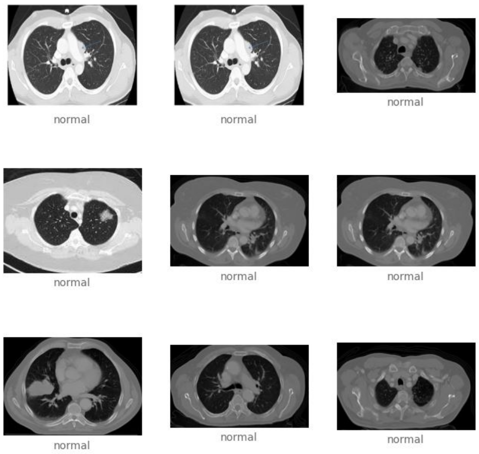


Figure 3.1: Normal cell

Normal cells are understood as healthy cells. It means that these cells are cancer free cells because there is no abnormality in them. Negative impacts like smoking (both active and passive), air pollution or other kinds of damages were not found analyzing them.

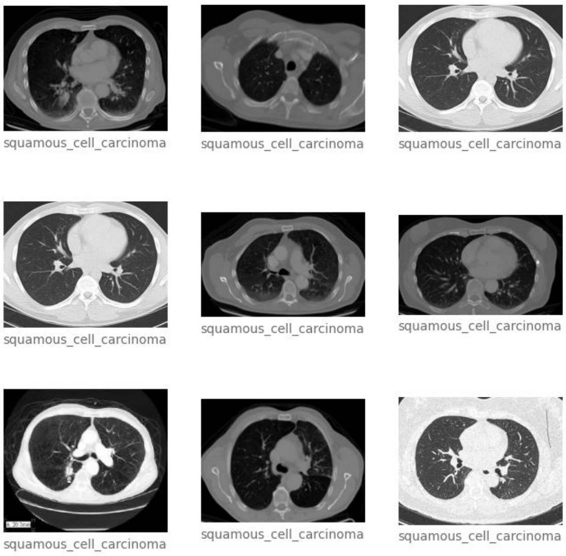


Figure 3.2: Squamous cell

Squamous cell carcinoma is a non small cell lung cancer illness, according to National Cancer Institute[15]. These kinds of tumors can be found in the middle section of the lung, the right or left side of bronchus also known as the main airway of the lung.

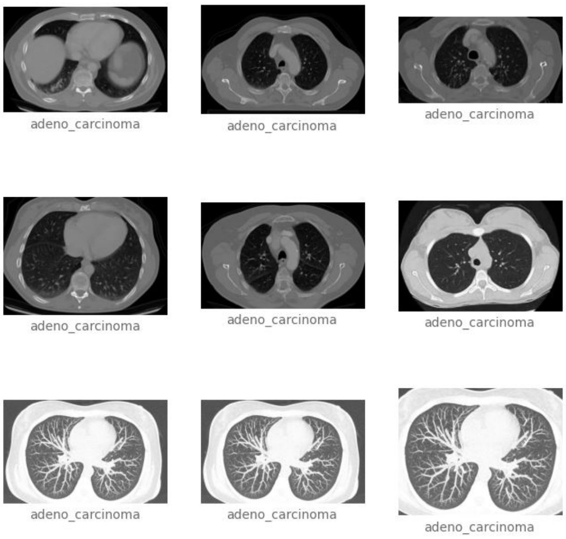


Figure 3.3: Adenocarcinoma affected cell

Outside the lungs, there is existence of cell linings. Sometimes they become cancerous. This kind of abnormality is known as adenocarcinoma. The American Society of Clinical Oncology [14] says that, nearly 40 percent of non-small cell lung cancers are occurred because of this.

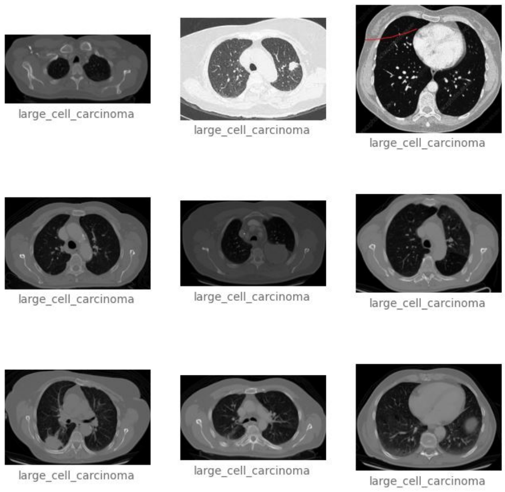


Figure 3.4: Large cell carcinoma

Large cell carcinoma is a very dangerous type of lung cancer as it grows and spreads faster than other lung cancers. It is another non-small cell lung cancer. It is not part of any specific subgroup[13].

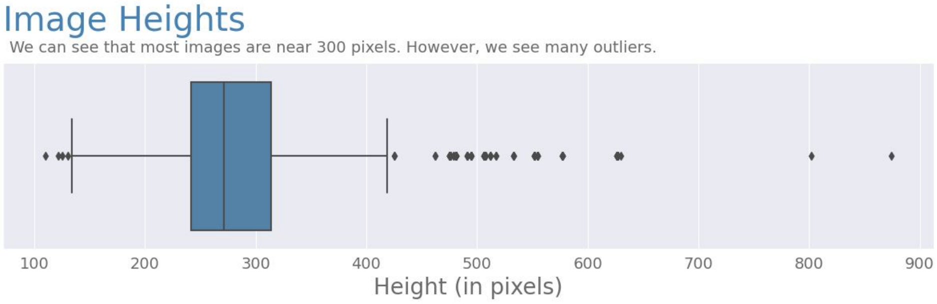


Figure 3.5: image height

The heights of the images were measured in pixels. Analyzing all the images, we came to an understanding that most of the images are around 300 pixels in height. However, there are outliers(not totally have similarity with the majority of items) also.

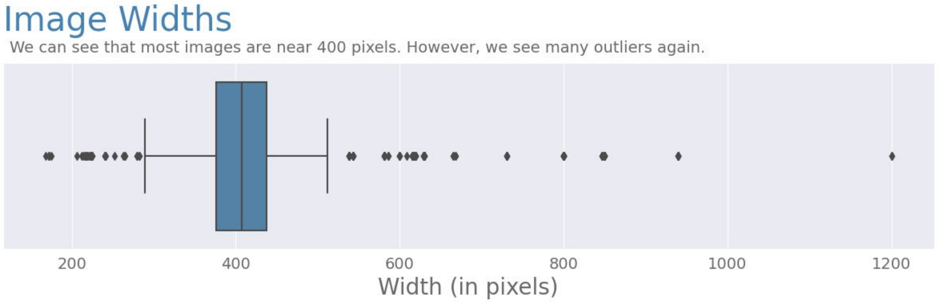


Figure 3.6: image width

The width of the pictures were also measured in the same unit as height, which is pixels. We learned after inspection that most images belong to category of 400 pixels height. Outliers were also spotted here.

3.2 Data Processing

Before starting the main procedure, the data has to be pre-processed. Before doing that, it was loaded into google colaboratory. The data was uploaded into google drive, it was mounted with colaboratory. Usually, a data set is a table of many rows and columns. This table has to be prepared to train or run by removing columns, or changing the item types etc. But ours is an image data set, so we had to approach differently. The data was visualized through coding and the images were counted for training data, testing data, valid data differently for different classes(adenocarcinoma, squamous etc). Image classification is necessary because it can decide whether a cell is affected or disease free[7]. The file types of data were also classified whether it is jpg or png.

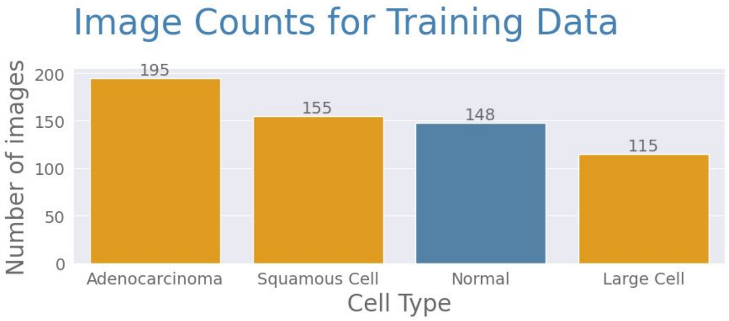


Figure 3.7: image count

The images had to be counted according to their classes. Each class (adenocarcinoma, normal or other cell) did not have same amount of images. They were classified well enough for future requirements. There is a picture of training data after classification with counting results.

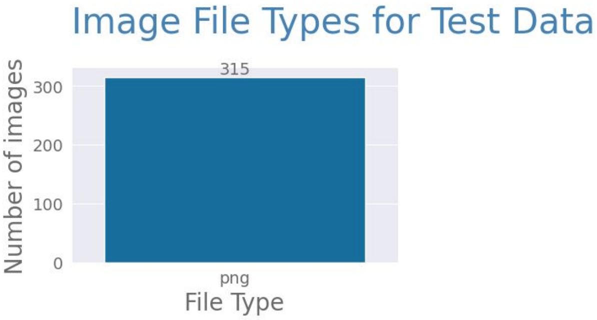
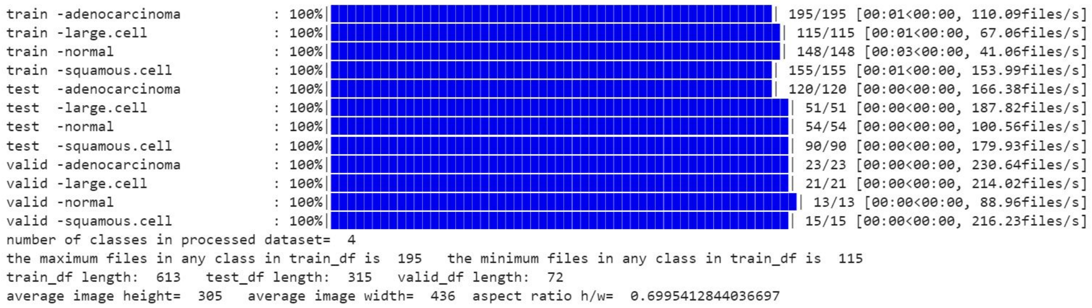


Figure 3.8: image type

As mentioned before, image file types were more than one. It was necessary to classify the image type also for taking input according to the file format. We can see a picture of PNG type images are in total 315 in the above figure.

3.3 Data Balancing

After processing, the data had to be balanced. Otherwise, the test results were becoming very one sided, as in biased. The class containing more data were being less benefited and class having lesser data were more favored. The models were finding it difficult to give a much appreciable result. Balanced data gave us higher accuracy. The data was augmented. The images were modified to increase the quantity of the data as larger data is more recommendable to work with. Some images were zoomed, cropped and then added to the data because items outside the data cannot be used. After augmentation, data inputs increased and each class had the same number of images indicating a balanced dataset.



```
train -adenocarcinoma          : 100%|██████████| 195/195 [00:01<00:00, 110.09files/s]
train -large.cell              : 100%|██████████| 115/115 [00:01<00:00, 67.06files/s]
train -normal                  : 100%|██████████| 148/148 [00:03<00:00, 41.06files/s]
train -squamous.cell           : 100%|██████████| 155/155 [00:01<00:00, 153.99files/s]
test  -adenocarcinoma          : 100%|██████████| 120/120 [00:00<00:00, 166.38files/s]
test  -large.cell              : 100%|██████████| 51/51 [00:00<00:00, 187.82files/s]
test  -normal                  : 100%|██████████| 54/54 [00:00<00:00, 100.56files/s]
test  -squamous.cell           : 100%|██████████| 90/90 [00:00<00:00, 179.93files/s]
valid -adenocarcinoma          : 100%|██████████| 23/23 [00:00<00:00, 230.64files/s]
valid -large.cell              : 100%|██████████| 21/21 [00:00<00:00, 214.02files/s]
valid -normal                  : 100%|██████████| 13/13 [00:00<00:00, 88.96files/s]
valid -squamous.cell           : 100%|██████████| 15/15 [00:00<00:00, 216.23files/s]
number of classes in processed dataset=  4
the maximum files in any class in train_df is  195   the minimum files in any class in train_df is  115
train_df length:  613   test_df length:  315   valid_df length:  72
average image height=  305   average image width=  436   aspect ratio h/w=  0.6995412844036697
```

Figure 3.9: data balancing

Chapter 4

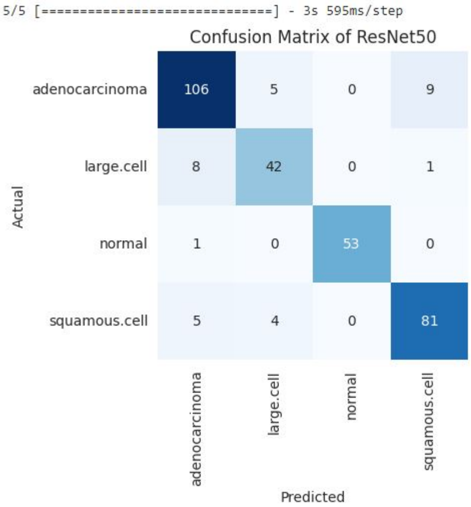
Result Analysis

4.1 Models Implementation

We used various CNN models to train our data set. We used 250 images for each class. First, we used models which are known and useful for these kinds of experiments. After that, we tried to make a model of our own and compare it with previously used models. Different models gave different accuracy as results.

4.1.1 ResNet-50

It is a 50 layers deep convolutional Neural Network. A pretrained version neural network on more than a million pictures can be loaded with this. This model has 5 stages. Each stage has a identity block and a convolution block. Both kinds of blocks have 3 convolution layers also. The model also has over 23 million parameters. It also has 15 channels. Among them, 3 channels are RGB and rest 12 channels are additional channels. It can take an input image of heavy height, width and channel width. So it has 3 dimensions and the input size is 224x224x3. Having many features and advanced characteristics, this model can achieve up to 97% accuracy. Resnet50 gave an accuracy of 89.52%. Which is very good. The accuracy and loss graph is given in Figure 4.2.



```
5/5 [==============================] - 3s 595ms/step
```

Confusion Matrix of ResNet50

Actual \ Predicted	adenocarcinoma	large.cell	normal	squamous.cell
adenocarcinoma	106	5	0	9
large.cell	8	42	0	1
normal	1	0	53	0
squamous.cell	5	4	0	81

Figure 4.1: resNet50 confusion matrix

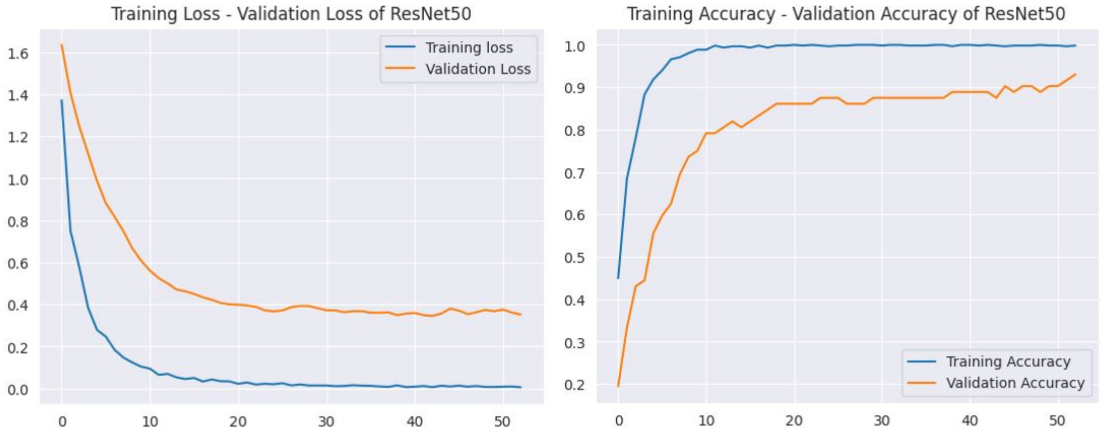
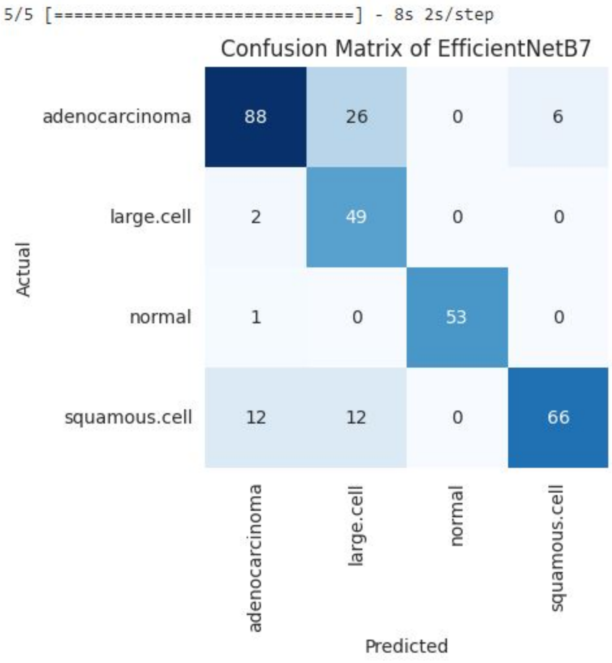


Figure 4.2: result resNet50

4.1.2 EfficientNetB7

it is a pure convolutional model, which is mobile friendly. It suggests a new method of scaling. It is known for its exceptional efficiency and accuracy in image recognition tasks. In this process, all the dimensions (width, depth, resolution) are scaled in a very simple but effective way. During this scaling process, a compound coefficient is used. This model has total 813 layers. The resolution of this model is 600X600 and its input size is 244x224x3. EfficientNetB7 gave accuracy of 81.27%. The accuracy and loss graph is given in Figure 4.4.



```
5/5 [==============================] - 8s 2s/step
```

Confusion Matrix of EfficientNetB7

	adenocarcinoma	large.cell	normal	squamous.cell
adenocarcinoma	88	26	0	6
large.cell	2	49	0	0
normal	1	0	53	0
squamous.cell	12	12	0	66

Figure 4.3: EfficientNetB7 confusion matrix

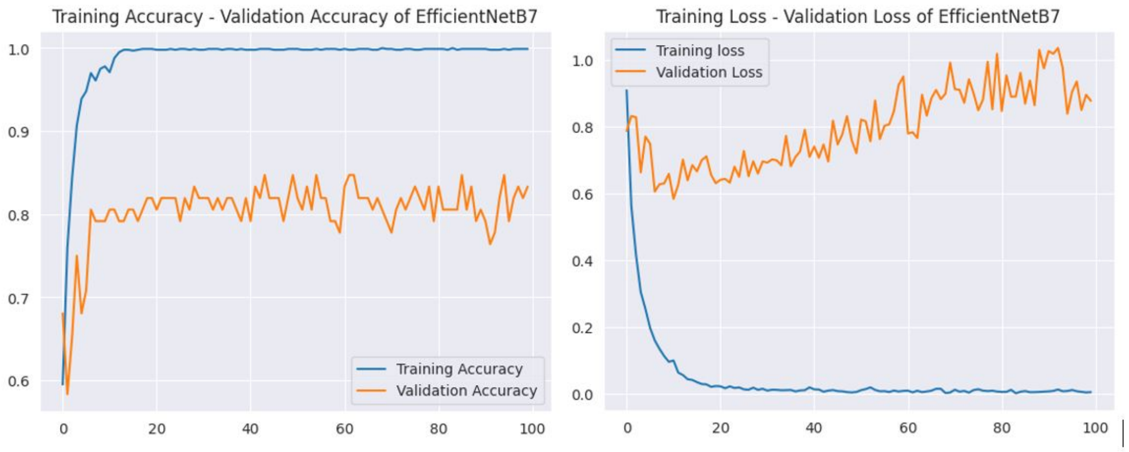
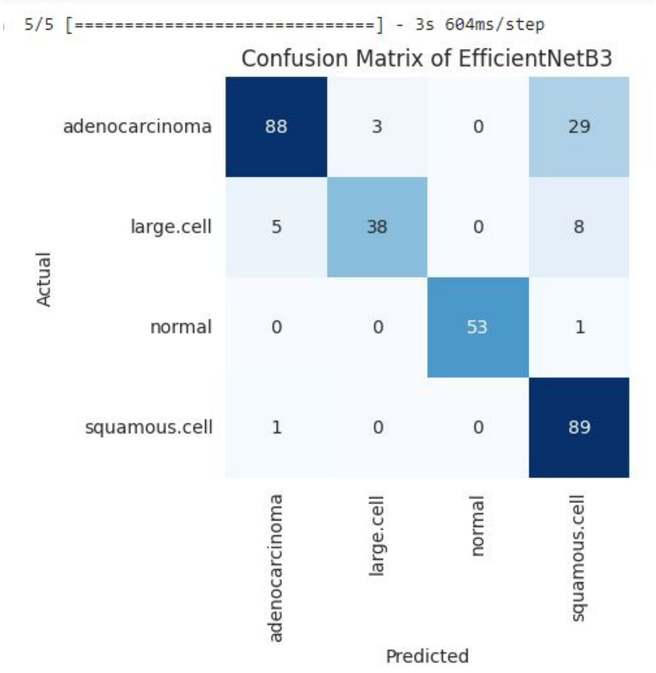


Figure 4.4: result EfficientNetb7

4.1.3 EfficientNetB3

Like B7, it is also a pure convolutional model, which is mobile friendly. It also suggests a new method of scaling. It is a widely used and effective model in computer vision due to its versatility and efficiency. It has consistently achieved impressive results in diverse image classification tasks, gaining popularity among researchers and practitioners in the field. EfficientNet models are better than ResNet models because if reasonable parameters are used, then top level results can be achieved. Its input size is also 244x224x3 and the resolution is 300X300. EfficientNetB3 gave accuracy of 85.08%. The accuracy and loss graph is given in Figure 4.6.



```
5/5 [==============================] - 3s 604ms/step
```

Confusion Matrix of EfficientNetB3

Actual \ Predicted	adenocarcinoma	large.cell	normal	squamous.cell
adenocarcinoma	88	3	0	29
large.cell	5	38	0	8
normal	0	0	53	1
squamous.cell	1	0	0	89

Figure 4.5: EfficientNetB3 confusion matrix

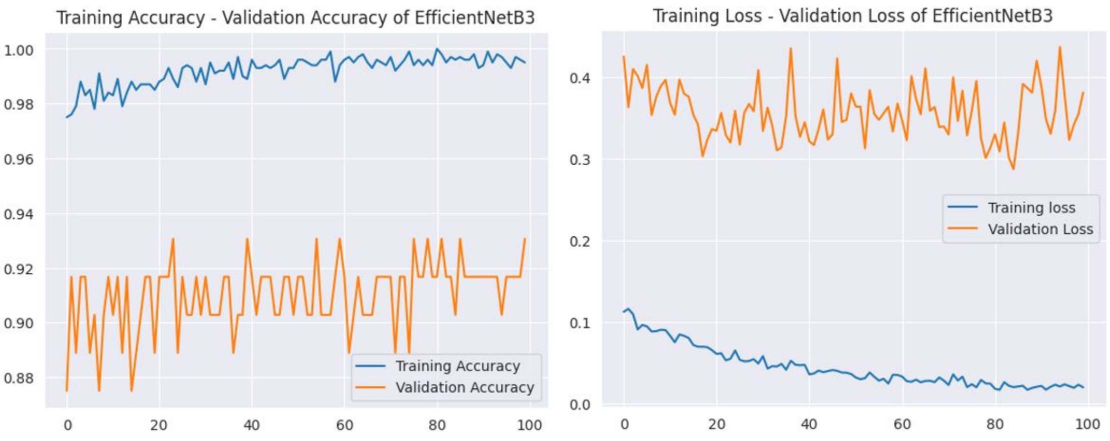
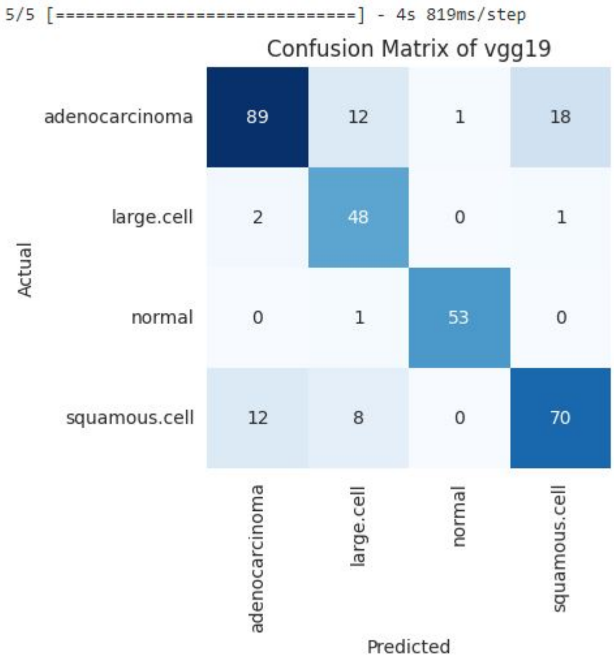


Figure 4.6: result efficientNetb3

4.1.4 VGG19

Another very deep layered convolution neural network is VGG-19. There are 19 layers altogether. The pretrained network can categorize pictures into 1000 object categories like living organs, input devices, etc. For a variety of photos, the network has taught rich feature representations. The network accepts images with a resolution of 224 by 224. Previously processed inputs of VGG19 will convert the RGB photos to BGR, then zero-center each color channel without scaling in accordance with the ImageNet dataset. Overall, this model can provide improved accuracy and is simple to use. It gave accuracy of 82.54



```
5/5 [==============================] - 4s 819ms/step
```

Confusion Matrix of vgg19

Actual	adenocarcinoma	large.cell	normal	squamous.cell
adenocarcinoma	89	12	1	18
large.cell	2	48	0	1
normal	0	1	53	0
squamous.cell	12	8	0	70

Figure 4.7: VGG19 confusion matrix

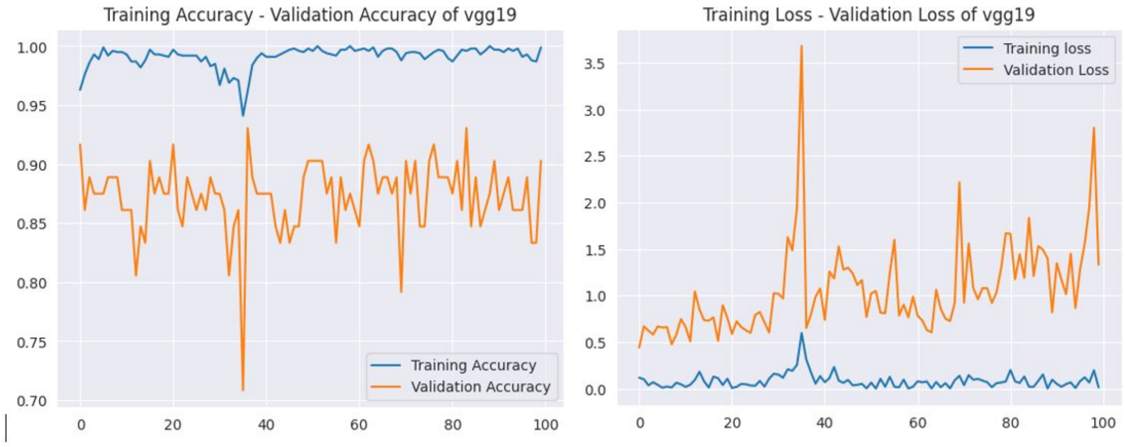
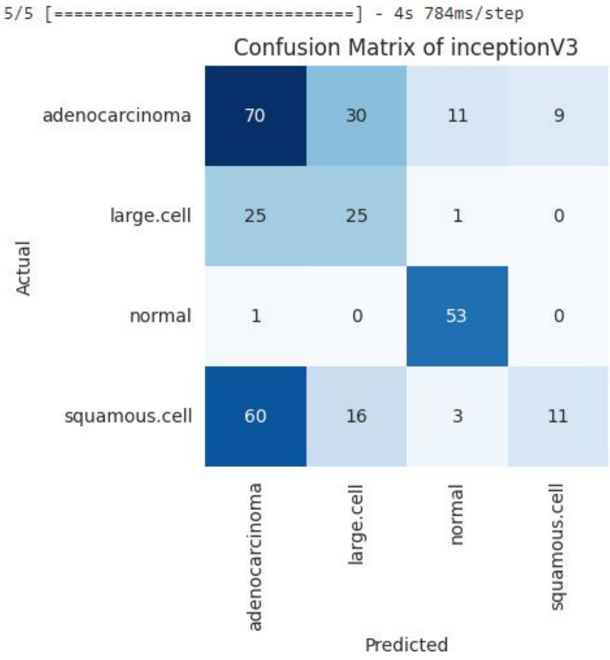


Figure 4.8: result VGG19

4.1.5 Inception V3

InceptionV3 is a very popular widely used image classified model. It is 48 layered deep. It has already used for more than one million images. The key innovation of InceptionV3 lies in its inception modules, which employ multiple parallel convolutional filters of different sizes, allowing the network to capture and integrate information at various scales. It's an advanced model of inception V1 model.It has a better grid side reduction and separate the convolution into smaller convolutions. This model is not that much expensive. It uses auxiliary classifiers as regularizes. It has a better network compared to the Inception V1 and V2 models, but its speed isn't compromised. The accuracy and loss graph is given in Figure 4.10.



```
5/5 [==============================] - 4s 784ms/step
```

Figure 4.9: InceptionV3 confusion matrix

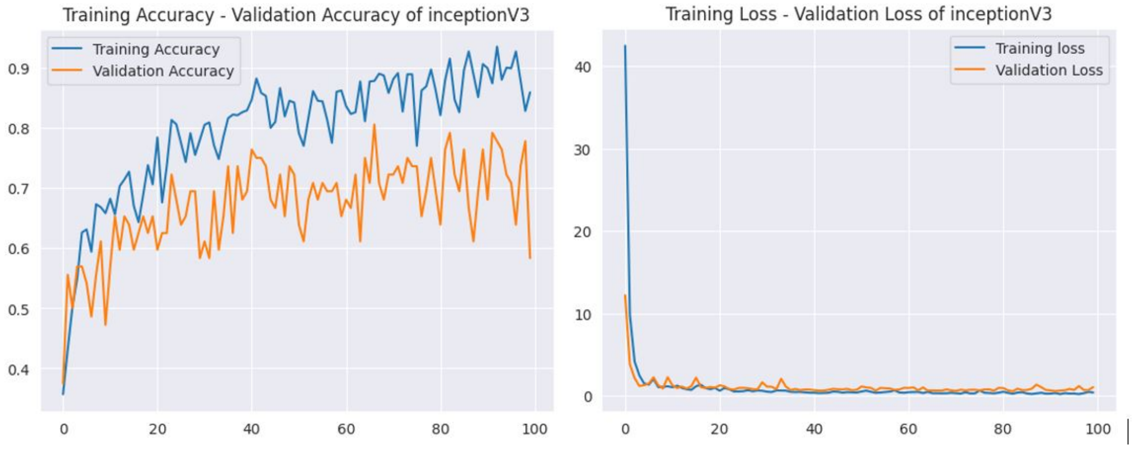
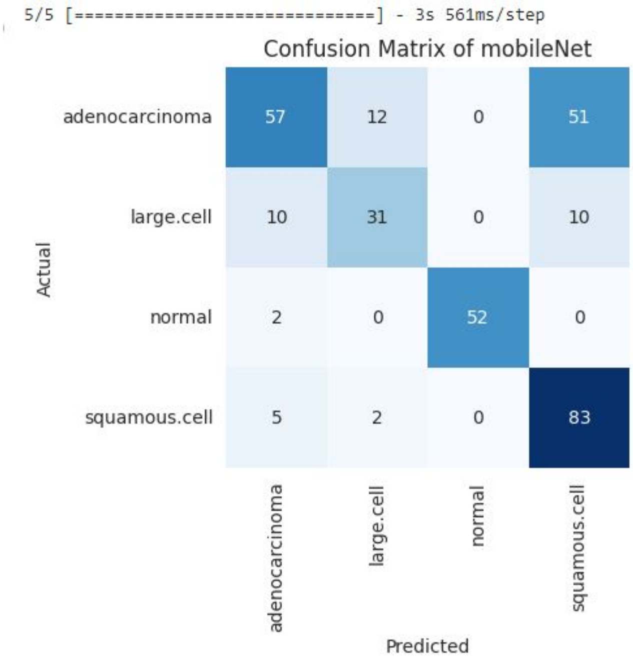


Figure 4.10: result InceptionV3

4.1.6 MobileNet

A network architecture defined as MobileNet uses depth-wise separable convolution as its fundamental building block. The depth-wise separable convolution in this model contains two layers: 3D point convolution The quantity of multiply-accumulates (MACS) determines the network's speed and energy consumption. That can help figure out how many addition and multiplication operations were merged. The models are tiny, low-latency, and low-power in order to fulfill the resource limitations of various types of use cases. It is employed in segmentation, detecting embedding, and classification. Its accuracy result is 50.48%. The accuracy and loss graph is given in Figure 4.12.



```
5/5 [==============================] - 3s 561ms/step
```

Figure 4.11: MobileNet confusion matrix

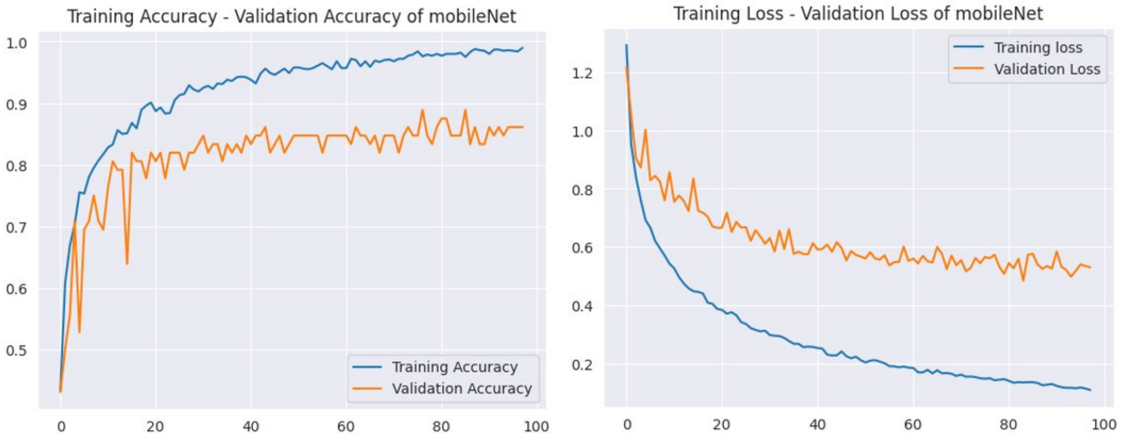
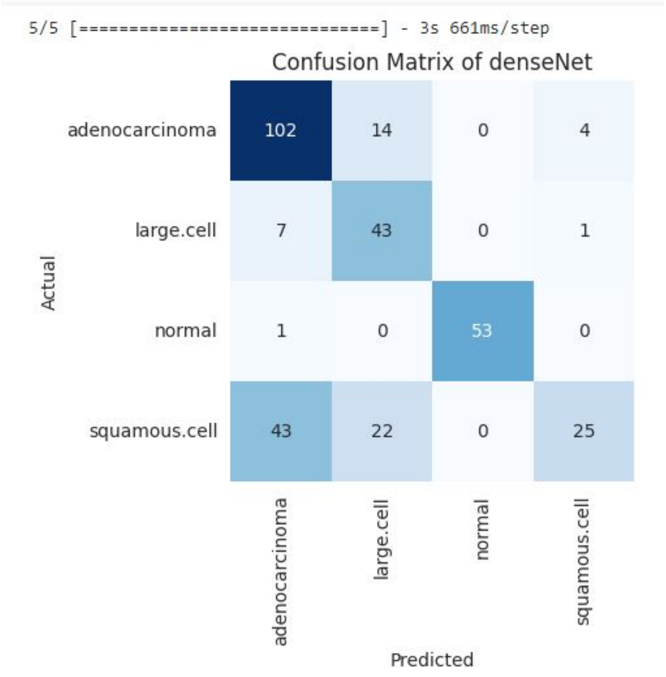


Figure 4.12: result MobileNet

4.1.7 DenseNet

DenseNet is a deep learning architecture. Its dense connectivity pattern and exceptional performance in image classification tasks are very good. It can improve the declined accuracy occurred by faded slopes by the upper level neural networks. It establishes direct connections between all layers, facilitating rich feature reuse and a seamless information flow. It can vanish gradients and can extract diverse and discriminative feature. Leveraging both dense and skip connections, it achieves outstanding accuracy with fewer parameters, making it a favored choice in computer vision applications like object recognition, segmentation, and image generation. It can strengthen the feature propagation, reduce the quantity of parameters. Features can be reused in this model [1]. This model gave the accuracy of 70.79%. The accuracy and loss graph is given in Figure 4.14.



```
5/5 [==============================] - 3s 661ms/step
```

Confusion Matrix of denseNet

Actual \ Predicted	adenocarcinoma	large.cell	normal	squamous.cell
adenocarcinoma	102	14	0	4
large.cell	7	43	0	1
normal	1	0	53	0
squamous.cell	43	22	0	25

Figure 4.13: DenseNet confusion matrix

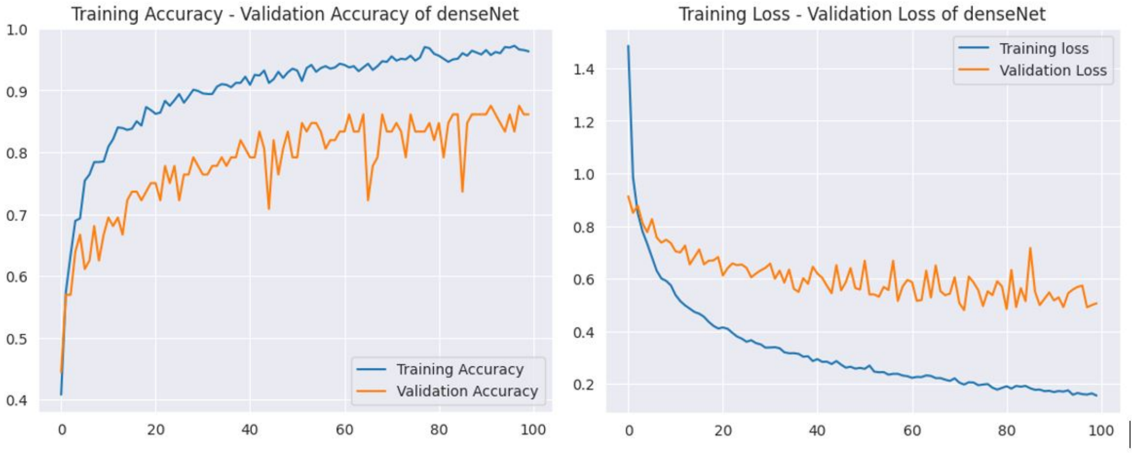
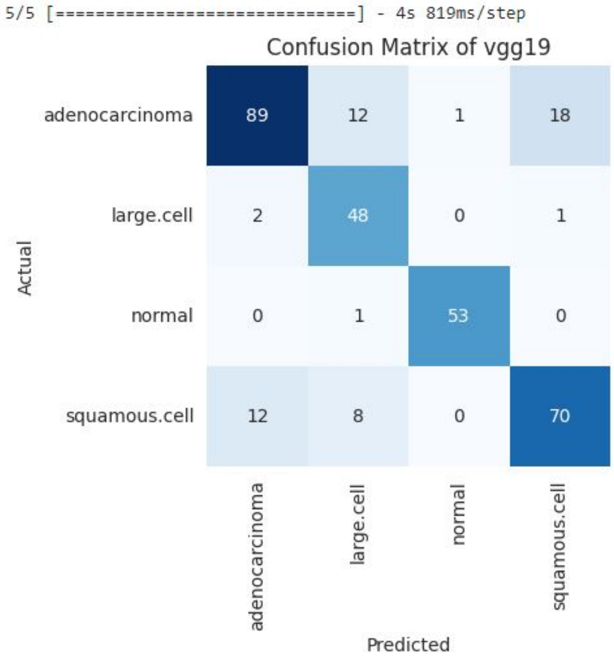


Figure 4.14: result DenseNet

4.1.8 Custom Model

We made an effort to suggest hyper parameter tuning in our custom model. It is the process of choosing the hyperparameters' ideal values in a machine learning model. Hyperparameters are variables that are chosen before the learning process starts and cannot be identified by data independently. This method was used as it plays a pivotal role in maximizing the performance of machine learning models, ensuring their ability to generalize well, improving stability, and enabling adaptability to diverse datasets. They define the structure and behavior of the model and have a significant impact on its performance. Unfortunately, the model could not perform well as there was not enough time. It gave accuracy of 57.78%. The accuracy and loss graph is given in Figure 4.16.



```
5/5 [==============================] - 4s 819ms/step
```

Confusion Matrix of vgg19

Actual \ Predicted	adenocarcinoma	large.cell	normal	squamous.cell
adenocarcinoma	89	12	1	18
large.cell	2	48	0	1
normal	0	1	53	0
squamous.cell	12	8	0	70

Figure 4.15: Custom model confusion matrix

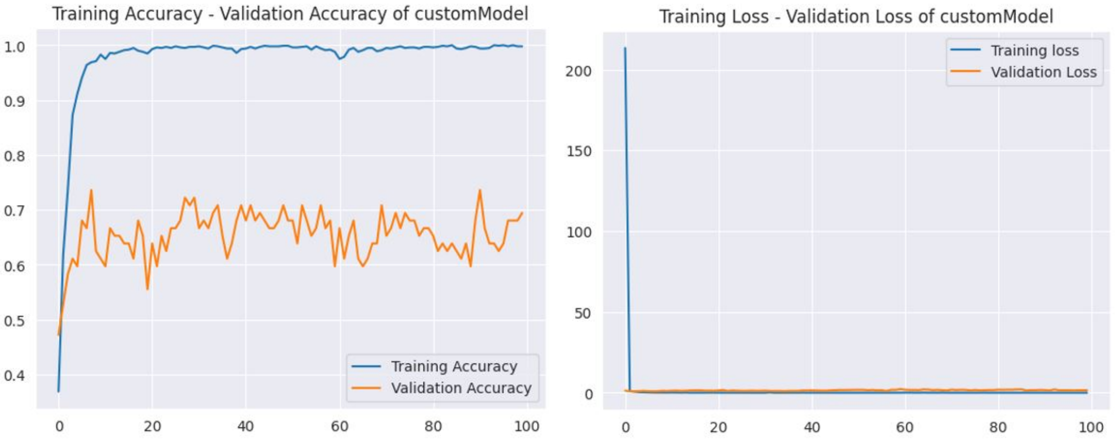


Figure 4.16: result Custom Model

4.2 Results in a Table

The table below shows the accuracy, error rate, f1 score, precision and recall for each model respectively.

Model	Accuracy	Error Rate	F1 score	Precision	Recall
resNet50	89.52%	10.48%	89.54%	89.56%	89.52%
efficientNetB7	81.27%	18.73%	81.83%	85.00%	81.27%
efficientNetB3	85.08%	14.92%	85.12%	87.83%	85.08%
VGG19	82.54%	17.46%	82.53%	83.48%	82.54%
denseNet	70.79%	29.21%	68.06%	75.16%	70.79%
inceptionV3	50.48%	49.52%	46.57%	51.87%	50.48%
mobileNet	70.79%	29.21%	69.93%	74.11%	70.79%
customModel	57.78%	42.22%	57.95%	59.46%	57.78%

Table 4.1: Overall report

Chapter 5

Conclusion

This experiment was an effort to add some advantage in detecting lung cancer cells in a human body as this illness is troubling a lot of lives for many years. We hoped that our work is going to be very useful for mankind and we tried to propose a model to help solving this problem. It is a matter of sorrow that we could not achieve the result we expected as there was a shortage of time running the proposed model. But, there is a high chance to get a far better result given the fair amount of time. So, we are optimistic that we can do better with our research and also extend it in the future. Given any opportunity, we will try to help the world curing lung cancer.

Bibliography

[1] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[2] B. B. Nair, K. Anju, and A. Jeevakumar, “Tobacco smoking induced lung cancer prediction by lc-micrornas secondary structure prediction and target comparison,” in *2017 2nd International Conference for Convergence in Technology (I2CT)*, IEEE, 2017, pp. 854–857.

[3] Q. Wu and W. Zhao, “Small-cell lung cancer detection using a supervised machine learning algorithm,” in *2017 international symposium on computer science and intelligent controls (ISCSIC)*, IEEE, 2017, pp. 88–91.

[4] J. Alam, S. Alam, and A. Hossan, “Multi-stage lung cancer detection and prediction using multi-class svm classifie,” in *2018 International conference on computer, communication, chemical, material and electronic engineering (IC4ME2)*, IEEE, 2018, pp. 1–4.

[5] W. Rahane, H. Dalvi, Y. Magar, A. Kalane, and S. Jondhale, “Lung cancer detection using image processing and machine learning healthcare,” in *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*, IEEE, 2018, pp. 1–5.

[6] Ö. Günaydin, M. Günay, and Ö. Şengel, “Comparison of lung cancer detection algorithms,” in *2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)*, IEEE, 2019, pp. 1–4.

[7] K. L. P. K. Balaji ME, *Medical Image Analysis With Deep Neural Networks*, https://www.sciencedirect.com/topics/engineering/image-classification, Accessed: 2023-05-22, 2019.

[8] P. Radhika, R. A. Nair, and G. Veena, “A comparative study of lung cancer detection using machine learning algorithms,” in *2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, IEEE, 2019, pp. 1–4.

[9] S. Mukherjee and S. Bohra, “Lung cancer disease diagnosis using machine learning approach,” in *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, IEEE, 2020, pp. 207–211.

[10] 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)*, IEEE, 2020, pp. 108–115.

[11] D. M. Abdullah, A. M. Abdulazeez, and A. B. Sallow, "Lung cancer prediction and classification based on correlation selection method using machine learning techniques," *Qubahan Academic Journal*, vol. 1, no. 2, pp. 141–149, 2021.

[12] M. Bala, V. Athira, and A. Rajendran, "Efficient multi-level lung cancer prediction model using support vector machine classifier," *IOP Conference Series: Materials Science and Engineering*, vol. 1012, p. 012 034, Jan. 2021. DOI: 10.1088/1757-899X/1012/1/012034.

[13] R. Zimlich, *Healthline*, https://www.healthline.com/health/lung-cancer/large-cell-carcinoma, Accessed: 2023-5-22, 2021.

[14] P. Baik, *Cancer center*, https://www.cancercenter.com/cancer-types/lung-cancer/types/adenocarcinoma-of-the-lung, Accessed: 2023-5-22, 2023.

[15] NCI, *National cancer institute*, https://www.cancer.gov/publications/dictionaries/cancer-terms/def/squamous-cell-carcinoma-of-the-skin, Accessed: 2023-5-22, N.A.