DEPARTMENT OF COMPUTER SCIENCE ENGINEERING,

SCHOOL OF ENGINEERING AND TECHNOLOGY,

SHARDA UNIVERSITY, GREATER NOIDA

A CNN based Sub-Classification of Non-Small Cell Lung Cancer into Squamous Cell Carcinoma and Adenocarcinoma

A project submitted

In partial fulfillment of the requirements for the degree of

Bachelor of Technology in Computer Science and Engineering

by

PRINCE BHARDWAJ (2018013267)

PRIYAM RASTOGI (2018004047)

LALAWMPUIA (2019003189)

RASIK DEV RATHOUR (2018004438)

Supervised by:

Dr. Rajendra Kumar, Assoc. Prof (CSE)

May 2022

DECLARATION

We, Prince Bhardwaj, Priyam Rastogi, Lalawmpuia, and Rasik Dev Rathour, hereby declare that the project work entitled "**A CNN based Sub-Classification of Non-Small Cell Lung Cancer into Squamous Cell Carcinoma and Adenocarcinoma**" submitted to Sharda University, is a record of an original work done by us under the guidance of Dr. Rajendra Kumar, Assoc. Prof(CSE).

This project work is submitted in the partial fulfillment of the requirements for the award of the degree of "Bachelor of Technology", in Computer Science & Engineering. The results embodied in this report have not been submitted to any other University or Institute for the award of any degree or diploma.

Name and signature of Students
PRINCE BHARDWAJ (2018013267)
PRIYAM RASTOGI (2018004047)
LALAWMPUIA (2019003189)
RASIK DEV RATHOUR (2018004438)

CERTIFICATE

This is to certify that the report entitled "A CNN based Sub-Classification of Non-Small Cell Lung Cancer into Squamous Cell Carcinoma and Adenocarcinoma" submitted by "PRINCE BHARDWAJ (2018013267), PRIYAM RASTOGI (2018004047), LALAWMPUIA (2019003189) and RASIK DEV RATHOUR (2018004438)" to Sharda University, towards the fulfillment of requirements of the degree of "Bachelor of Technology", is a record of bonafide final year project work carried out by them in the "Department of Computer Science and Engineering, School of Engineering and Technology, Sharda University".

The results/findings contained in this project have not been submitted in part or whole to any other University/Institute for the award of any other Degree/Diploma.

Name: Dr. Rajendra Kumar
Supervisor and Assoc. Prof (CSE)

Name: Prof. (Dr.) Nitin Rakesh
Head of Department
Place: Sharda University
Date: 13-05-2022

Signature of External Examiner

Date: 13-05-2022

ACKNOWLEDGEMENT

A major project is a golden opportunity for learning and self-development. We consider ourselves very lucky and honored to have so many wonderful people lead us through in completion of this project.

First and foremost, we would like to thank Dr. Nitin Rakesh, HOD, CSE, who gave us an opportunity to undertake this project.

Our thanks to **Dr. Rajendra Kumar** for his guidance in our project work. He, in spite of being extraordinarily busy with academics, took time out to hear, guide, and keep us on the correct path. We do not know where we would have been without his help.

The CSE department monitored our progress and arranged all facilities to make life easier. We choose this moment to acknowledge their contribution gratefully.

PRINCE BHARDWAJ (2018013267)

PRIYAM RASTOGI (2018004047)

LALAWMPUIA (2019003189)

RASIK DEV RATHOUR (2018004438)

ABSTRACT

Lung cancer is an important healthcare concern. It affects gender in different ways. Men are at the greatest risk of dying from this cancer, and women are at the second-highest risk. The main objective of this work is to classify between squamous cell carcinoma and adenocarcinoma. Around 85 percent of all lung cancer cases are non-small cell lung cancer. Early detection and treatment are important to a patients' recovery. Diagnosing the various kinds of cancers of the lungs is usually a troublesome process that requires time and error. In addition to identifying lung cancer subtypes more accurately and in less time, convolutional neural networks may help in determining patients' right treatment procedures and their survival rates. Despite its complexity, even for experienced pathologists, this area of research can be challenging when it comes to adenocarcinoma and squamous cell carcinoma. In this project, we proposed a system for sub-classification of non-small cell lung cancer into squamous cell carcinoma and adenocarcinoma. The proposed system is able to classify non-small cell lung cancer into squamous cell carcinoma and adenocarcinoma.

Index Terms- Identification, Classification, Deep Learning, Machine learning, Convolution neural network, Image processing, Accuracy, Confusion matrix

Contents

Chapter 1: INTRODUCTION

Cancer is now becoming the most common disease and Lung cancer is the major cause of cancer death in the world. It is difficult to identify because it appears and manifests symptoms only at the end of the process. Early discovery and treatment of the condition, on the other hand, can minimize the mortality rate and probability. Approximately 7.6 million people worldwide die each year from lung cancer, as per reports by the World Health Organization. Furthermore, the number of people dying from cancer is expected to rise further, reaching roughly 17 million by 2030. To diagnose lung cancer, there are different methods available, such as MRI (magnetic resonance imaging), CT (Computed tomography), isotopes, and X-rays. The specialists use appropriate tools to analyze images and detect and diagnose lung cancer at various levels of severity. Different laboratory and clinical approaches are used, such as chemical treatments to destroy malignant cells or stop their replications, targeted therapies, and radiotherapy. Patients must endure painful, time-consuming, and expensive treatments in order to identify and detect cancer disorders. To identify and diagnose lung cancer, image processing tools and machine learning approaches were utilized. The aim is to make a CNN model which can recognize lung cancer and distinguish between squamous cell carcinoma and adenocarcinoma. The purpose of this project “Lung Cancer Detection System” is to help and assist doctors and medical actors in detecting and classifying what type of a Lung Cancer is with the use of a deep learning technique as it is one of the best model for medical image processing, lung nodule detection and classification, detection of lung cancer and feature extraction.

1.1 Problem Definition

Around 85 percent of all lung cancer cases are non-small cell lung cancer. Early detection and treatment are important to a patients’ recovery. Diagnosing the various kinds of cancers of the lungs is usually a troublesome process that requires time and error. In addition to identifying lung cancer subtypes more accurately and in less time, convolutional neural networks may help in determining patients’ right treatment procedures and their survival rates. Despite its

complexity, even for experienced pathologists, this area of research can be challenging when it comes to adenocarcinoma and squamous cell carcinoma.

1.2 Project Overview/Requirement Specification

1.2.1 Functional Requirements

1.2.1.1 Introduction

In this project, we proposed a system for sub-classification of non-small cell lung cancer into squamous cell carcinoma and adenocarcinoma. The proposed system is able to classify non-small cell lung cancer into squamous cell carcinoma and adenocarcinoma. The system will provide a platform for the doctors to check the Lung Cancer image and differentiate between the two main types of cancer cells.

1.2.1.2 Input

Doctors should know how to use this system. Data which will be inputted for checking the lung cancer type must be correct.

1.2.2 Normal Requirements

These are the demands which must be present for customer satisfaction.

N1: The program should provide a user interface with graphics.

N2: System should identify correct type of lung cancer.

N3: Patients should receive their reports on time

1.2.3 Non functional Requirements

These are the specifications, as the name implies, that are not specifically correlated with particular functions offered by the device.

1.2.3.1 Performance Requirements

Execution based on the relation, high execution of the PC may involve one or more of the accompanying: fast reaction time for a certain bit of work. In comparison to the time and assets used, execution is defined by the measure of useful work done by a PC framework or PC system.

1.2.3.2 Reliability

Unwavering quality is a property of any component related to a PC (for example, programming, or equipment, or system) that performs consistently as its determinations indicate. For some time, it has been seen as one of three related characteristics that should be considered when making, buying, or using an object or part of a PC.

1.2.3.3 Availability

Accessibility is a general concept used in PC systems and system management to describe the measure of time over a one-year span that the framework assets are available in the wake of partial system disappointments. A structure with all its properties that is continually available is seen as fruitful.

1.2.3.4 Security

Security (or PC security) in registration is the technique to ensure that information placed on a PC can not be accessed or negotiated without approval by any person. Data encryption and passwords are the majority of PC efforts to develop security. Encryption of information is the interpretation of data into a structure that is indiscernible without a method of disentanglement. A watchword is a mystery word or phrase that gives a client access to a particular project or structure.

1.2.3.5 Maintainability

It is defined as the probability within a given time of conducting a successful repair operation. As such, practicality tests the straightforwardness and speed at which, after a disappointment

occurs, a system can be returned to operating status. Convenience is a trademark that is credited to a PC application in the event that it may be used rather than the one in which it was developed as part of operating systems without the need for major reconstruction. Porting is the job of performing whatever work that is necessary to maintain the PC program going in the new environment.

1.2.3.6 Ability of Learning

It is simple to operate and reduces the learning function.

1.3 Hardware Specifications

Minimum Requirements	Windows
OS	Windows 7
Processor	Dual-core, Intel i3
RAM	8 GB RAM
DISK Space	Minimum 500GB of disk space is needed for this work

Table 1: Minimum Hardware Requirements of Project

1.4 Software Specifications

- Windows 7+
- Python 3.9 (64 Bit)
- Javascript
- Django
- Keras API
- Jupyter NoteBook
- VS Code

Chapter 2: LITERATURE SURVEY

2.1 Existing System

By referring to various survey papers we could find the following systems being in use. Here we will explain the literature survey that we conducted for the development of the project and refer to projects and systems that are existing and have been used in the market. These include:-

Suren Makaju et al. **[1]** proposed a method for Lung Cancer Detection Using CT Scan Images. By using watershed segmentation for detection and SVM for determining whether a nodule is malignant or benign, this system detects the cancerous nodule on lung CT scan images. The dataset that is used in this work is Image database consortium and this work got accuracy of 92%.

Asuntha et al. **[2]** presents deep learning for cancer detection and classification. Here, author used Fuzzy Particle Swarm Optimization (FPSO) for feature selection. The best feature extraction techniques were employed in this study, including Histogram of oriented gradients, Wavelet transform-based features, Local binary pattern, Scale invariant feature transform, and Zernike moment. This work used LIDC database and got 65.62% accuracy.

Atsushi Teramoto et al. **[3]** shows Automated Lung Cancer Type Classification Using Cytological Images (Automated Lung Cancer Type Classification Using Cytological Images) It's a Deep Convolutional Neural Network that's being employed. The author used a Deep Convolutional Neural Network to classify adenocarcinoma, squamous cell carcinoma, and small cell carcinoma. This job has a 71.1 percent accuracy rate.

Bijaya Kumar Hatuwal et al. **[4]** uses Convolutional Neural Network on Histopathological Images to presents Lung Cancer Detection. This work used CNN to classify the benign, adenocarcinoma, and squamous cell carcinoma.Here Author used LC2000 dataset & achieved training accuracy 96.11% and a validation accuracy of 97.20%.

Ruchita Tekade et al. **[5]** presents Lung cancer Detection and classification using Deep Learning. Here the author used CNN architecture and dataset used is Lung Image Database Consortium and LUNA16. The finding of this paper is combining the two approaches as proposed architecture and U-Net segmentation has given the better results for predicting lung nodule detection and also further predicting malignancy level. This approach gives accuracy as 95.66% and loss 0.09. The drawback of this paper is that the model showed slow detection rate.

Amjad Khan et al. **[6]** proposed a system for Identification of Lung Cancer Using Convolutional Neural Networks Based Classification. In this paper deep convolution neural network is used and LIDC dataset is used. In this work, The CT images of lung nodules have been classified as cancerous (malignant) or benign (benign) using deep convolutional neural networks. Therefore, preprocessing was performed before applying input CT images to the network work model to ensure that the images were of equal size and format. DCNN does not handle large-size images well, which is a drawback.

Nasrullah et al. **[7]** proposed Automated Lung Nodule Detection and Classification Using Deep Learning Combined with Multiple Strategies the team uses publically available datasets LUNA16 and LIDC-IDRI and hospital data were used for the training of nodule detection and classification, respectively.In the first step, we applied 3D Faster R-CNN with CMixNet and a U-Net-like encoder–decoder to determine the presence of nodules in the lung CT image. Further, In order to classify the nodules as benign or malignant, the 3D CMixNet with GBM was used to analyze the detected nodules. Last but not least, the deep learning-based nodule classification results were further analyzed by considering multiple factors, including patient family history, age, smoking history, clinical biomarkers, nodule location, and size. Several experiments have been run on the publically available LUNA16 and LIDC-IDRI datasets.

A. Asuntha et al. **[8]** presents Deep learning for lung Cancer detection and classification. The dataset used here is Aarthi Scan Hospital dataset. In this project Detecting the cancerous lung nodules from the given input lung image and to classify the lung cancer and its severity. To detect the location of the cancerous lung nodules, this work uses novel Deep learning methods. Here, features are classified using Deep learning. A novel FPSOCNN is proposed which reduces computational complexity of CNN. The drawback of this paper is visual interpretation or evaluation of a large number CT image slices remains as a difficult task.

Wafaa Alakwaa et al. **[9]** uses a 3D Convolutional Neural Network (3D-CNN) to identify and categorise lung cancer. The Kaggle Data Science Bowl (DSB) 2017 patient lung CT scan dataset was used in this study. This dataset provides labelled data for 1397 patients, separated into a 978-patient training set and a 419-patient test set. It also performs end-to-end CNN training from the ground up. The accuracy of the test set produced by this model was 86.6 percent.

In this paper Anum Masood et al. **[10]** will detect the Lung Cancer Detection and Classification by using the Automated Decision Support System via Enhanced RFCN with Multilayer Fusion RPN.The model has been trained and evaluated using LIDC dataset and clinical dataset. LIDCIDRI contains total of 1018 CT scans sets and every CT scan has around 300 slices.Each slice of which is gray level picture insize of 512*512 and the thickness of the slice is 3mm.The technique that was used by this model is Multidimensional Region-based Fully Convolutional Network.The accuracy that was achieved by this model is 97.91%.

Perumal et al. **[11]** detects and classifies lung cancer using CT scan pictures and better artificial bee colony optimization. The LIDC (7-9) dataset, developed by the National Cancer Institute (NCI) as part of the Lung Image Database Consortium, was used in this model. This model employs the region expand technique, the watershed algorithm, and the Artificial Bee Colony (ABC) algorithm. This model will be really accurate.

In this paper, Chao Zhang at al. **[12**] use a Deep Convolutional Neural Network to detect and classify lung cancer.This model includes three data sets for various reasons. The first group of CT scans came from the LUNA16 data set, the second set came from Guangdong Provincial People's Hospital's thoracic CT images,and the third group consisted of 50 individuals who had undergone surgical dissection. CNN is the algorithm that this model employs. The accuracy that is achieved by this model is 84.4%.

Nidhi S. Nadkarni et al. **[13]** is primarily concerned with the classification of lung images as normal or pathological. The median filter was employed to remove impulse noise from the images in their suggested technique.The use of mathematical morphological operations allows for reliable lung segmentation and tumour detection. Three geometrical features were recovered from the segmented region and supplied to the SVM classifier for classification: area, perimeter, and eccentricity.The accuracy that is achieved by this model is 86.6%.

In this paper, Meraj Begum Shaikh Ismail et al. **[14]** classify and detect lung cancer using the machine learning algorithms. K Means are used for segmentation. For biomedical picture segmentation, the U-Net Convolutional Network is used. It requires an input image and a region-of-interest output mask. The dataset that is used in this model is TCIA dataset, (LIDC-IDRI) and Kaggle data science bowl 2017. The features are taken from the segmented images, and classification is performed using a variety of machine learning algorithms. The accuracy, sensitivity, specificity, and classification time of the proposed methods are all measured. This model gives very accurate result.

Vaishnavi. D1 et al. **[15]** investigated a lung cancer detection algorithm in their study. They employed a discretely sampled wavelet in the Dual-tree complex wavelet transform (DTCWT) for pre-processing. GLCM is a second-order statistical method for texture analysis that tabulates how distinct Gray level combinations co-occur in a picture. It calculates the intensity variation at

the pixel of interest. They employed a Probability Neural Network (PNN) classifier that was tested for both training and classification accuracy. It provides accurate and quick classification.

In this paper, Prof. DR. K. Rajeswari et al. **[16]** investigated the notion of lung nodule detection and malignancy level prediction using lung CT scan images The datasets LUNA16, and Data Science Bowl 2017 were used in this experiment, which was run on a CUDA equipped GPU Tesla K20. The dataset was analysed using an Artificial Neural Network with the goal of extracting features and classification. They employed the U-NET design to segment lung nodules from CT scan images, as well as a 3D multigraph VGG-like architecture to identify lung nodules and forecast malignancy levels. The outcomes were better when these two approaches were combined. This method has a 95.66 percent accuracy, a loss of 0.09, a dice coefficient of 90 percent, and a 38 percent accuracy for predicting log loss.

2.2 Proposed System:

2.2.1 Dataset:

We discovered several commonly used datasets while conducting our survey:-

- Image database consortium
- LIDC dataset
- LC25000
- LUNA16

2.2.2 CNN(Convolutional Neural Network):-

This project is using a convolutional neural network for image classification. Because of its capacity to extract features from images without requiring complex pre-processing, Convolutional Neural Networks approaches are used.A convolutional layer, a pooling layer, and a fully linked layer make up CNN. The CNN's main building block is the convolution layer. It is responsible for the majority of the network's computational load. The pooling layer uses a summary statistic of neighbouring outputs to replace the network's output at specific spots. This

reduces the representation's spatial size, which reduces the amount of computation and weights necessary. Convolutional Neural Network (CNN) is a Deep Learning algorithm that can take an image as input, assign significance to various aspects/objects in the image, and distinguish one from the other. CNNs are widely used in the image recognition, image classification, object detection, recognition of faces etc. CNN image classification takes an input image and then process the image and then classify the image under certain categories. CNNs are another type of neural network that can be used for assisting machines in visualizing things and performing tasks such as object detection and classification. CNN is a specialized type of neural network whose model are designed for working with image data.As in a conventional fully convolutional neural network, fully connected layers have full connection with all neurons in the preceding and subsequent layers.

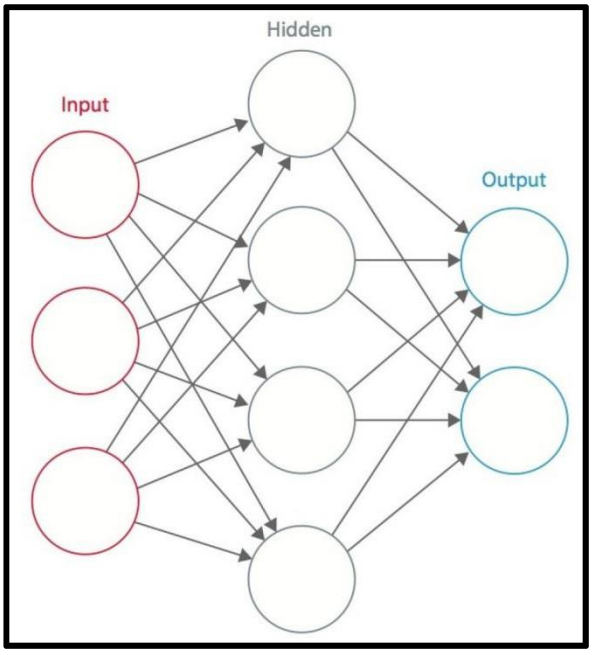

Fig 1: CNN

2.2.3 Data Preprocessing:-

In image pre-processing, the images in our dataset are cropped to the size of 299x299 pixel from 768x768 pixel to reduce the background area and 299x299 size is suitable for CNN method. Then the cropped image is used for training and validation dataset on 3-layer CNN. During image pre-processing author set the value of rescale is 1/255, shear_range is 0.20, zoom_range is 0.20 and validation_split is 0.20.

2.2.4 Image Preprocessing :-

For successful picture prediction, image preparation is essential in the identification pipeline. For accurate character prediction, image preprocessing is crucial in the identification pipeline. Noise reduction, image segmentation, cropping, and scaling are some of the other techniques used. These methods were mostly used in our experiment to classify our images.

2.3 Feasibility Study:

Lung cancer is an important healthcare concern. It is very dangerous for human's life. Early detection and treatment are important to a patients' recovery. Diagnosing the various kinds of cancers of the lungs is usually a troublesome process that requires time and error.

Here we are with an idea to build a web-based application through which the process of lung cancer detection becomes easy.

The user interface will be simple and easy to use by a common man. This application can be used by doctors as well as patients. It is used in hospital. Doctor will do a MRI (Magnetic resonance imaging) of the patient and get the image from that. Then this image is loaded in our model and then our model will able to classify between types of lung cancer i.e. Squanomous cell carcinoma and Adenocarcinoma images.

Doctor will send the report to patient account and the patient is also able to access the report by logging in the system.

- It is user-friendly web-based application.
- With respect to the healthcare sector, this application can be used by many hospitals to get accurate results of lung cancer types.
- It is compatible with a normal desktop/laptop which has a web browser installed in it and has internet access.
- The workload of the doctor can be reduced as they get fast and accurate results.
- It is economically affordable as it does not require any special requirements.
- This application will be user-friendly since the user interface will be simple and easy to use by any non-technical person.

For a feasibility analysis, any understanding of the scheme's major specifications is required. Computer's feasibility dimensions would be as indicated in

- **Finance**

  Is it feasible from a budgetary standpoint?

  Is it conceivable for the software company and its customer or company to reach a fair level of production?

- **Technology**

  Is the project feasible from a technological standpoint?

  Is it a part of the current state of the art?

  Will the failure be restricted to the requirement of a level implementation?

- **Time**

  Is it possible to beat the competition in terms of selling an idea?

- **Resources**

  Will the corporation have the capital necessary for success?

  Two major variables used in the study of viability are

  a) Technological Feasibility

  b) Cost Feasibility

2.3.1 Technical Feasibility

The purpose of this analysis is to check the technological viability, that is to say, the system's technical requirements. Any built system does not have a strong need for the technological resources required. This will add to intense strains on the intellectual resources available. It would bring to the customer's already firm hopes. Since this system can only be applied with minor to no modifications, a bare minimum must be met.

In the following ways, a practical evaluation of feasibility can be carried out.

i. NP-Complete

ii. NP-Hard

iii. Satisfiability

i. NP-Complete

The P Class comprises those issues which can be solved in polynomial time. The NP class consists of such concerns which are verified in polynomial time. If any other issue in NP can be converted (or reduced) into p in polynomial time, a question p in NP is NP-complete.

ii. NP-Hard

There are problems where no such viable solutions have been identified. The complexities of these topics are usually more complex unlike P, NP, and NP-Complete. Relatively high multiplicative constants, exponent terms, or polynomials of a high order can be involved in this.

iii. Satisfiability

In order to make it valid, if there is at least one way to add value to its vector and we denote it by using SAT, the Boolean formula is satisfied. The dilemma of evaluating whether or not the given formula is satisfactory.

2.3.2 Cost Feasibility

This study evaluates the economic impact of the scheme on the business. It restricts the amount of money that can spend on the research and development of its strategy. It is necessary to justify the expenses. Thus, within the budget, the developed system was also developed and this was done because much of the technology used is readily accessible. It was only appropriate to buy the personalized items.

2.4 Risk Management

2.4.1 Risk Identification

a. Product Size Related

R1 Memory may be squandered as a result of additional lines of code or redundant algorithms.

b. Customer Related

R2 Since its consumer isn’t a professional individual and it poses a challenge in interpreting the customer's additional specifications.

R3 If the consumer offers unnecessary details; it can result in an undisclosed danger.

c. Process Risk

R4 A wrong or unknown input image can result in an unexpected result.

d. Technical Risk

R5 The learning time would increase if the database would be large.

e. Development Environment Related

R6 When a client requests a replacement or asks for an unnecessary alteration after in the implementation step, it is impossible to change the whole system configuration to accommodate the request.

R7 Inexperience and a lack of tool preparation can make it challenging to complete project modules.

2.4.2 Strategies used to manage Risks

S1 By reducing redundant coding, we can prevent Chance R1.

S2 Meeting with the customer regularly reduces the risk to some extent.

S3 R3 properly develops the system to incorporate modifications at a later stage and retains all necessary paperwork to minimize the risk, as previously stated.

S4 Using a proper error handling technique can reduce the chance of getting unexpected results.

S5 Extracting the necessary features only from the raw data and storing them in the real-time database.

S6 As consumer demand changes, we will continue to increase the software's functionality.

S7 We will prevent R7 by providing adequate tool instruction.

Chapter 3: SYSTEM ANALYSIS & DESIGN

3.1 Software Requirement Specification

1. Frontend:

The frontend of the website is working on HTML5, and Django. User can use Chrome/ Safari/ Mozilla, any of the above browsers to run the website. The website is made in a way that each browser will support it, and its components will not break down at any time.

2. Backend:

The backend of the website is working on Django. Django is used to collect the data from the model and pass the data to the frontend.

3. Code Editor:

We have used three different code editors(VS Code/ Jupyter Notebook) for the development of the project.

i. Vs Code: VS Code editor is one of the most useful code editor for the development of the website. VS Code provides a lot of tools to make the development process easier. The auto formatting and extension of any language feature of the VS Code helps to avoid the syntax errors.

ii. Jupyter Notebook: Jupyter notebook is the library of python, which is very helpful to run and save the data sessionally. It is very helpful while working with the machine learning algorithms, as it will save the model locally, we need not to train the model again and again. This helps to save the time and memory utilization of the CPU.

FrontEnd	*Django, HTML, CSS*
BackEnd	*Python, Machine Learning*
Database	*sqlite3*
Code Editor	*Jupyter Notebook, VS Code*

Table 2: Software required for Project

3.2 : Flowcharts/USDs/ERDs

3.2.1 Flow chart

As the model is executed and run the UI is deployed to a Web app , here we have a login screen where we will have to provide our credentials in order to log inn to the Web App after logging inn we will redirected to a page where we can input image(Coloured image,Xrays etc) which shows what type of Lung Cancer it is. The uploaded image is then separated into smaller segments and CNN is applied to each of the segments. CNN extracts the percent of each type of cancer cell that are present in each segment and will Classify between which type of lung it is and it will give the output.

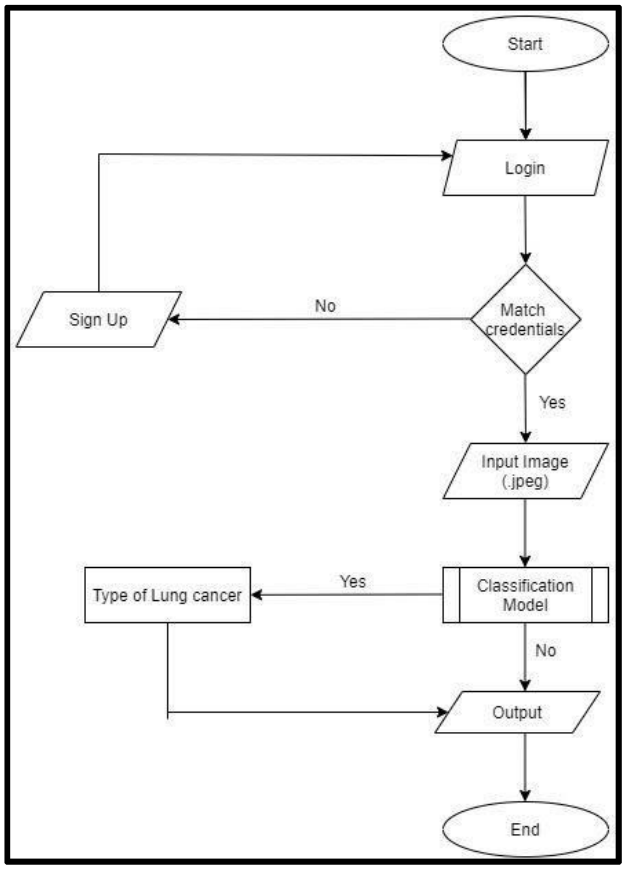

Fig 2: Flow Chart

3.2.2 Use case Diagram

Here a Doctor will will login with the his/her Credentials and will upload the Xray,coloured image etc and classify the image to get the output on what type of Lung Cancer it is.

Admin will work on maintaining the databases such data reporting, handling error etc.

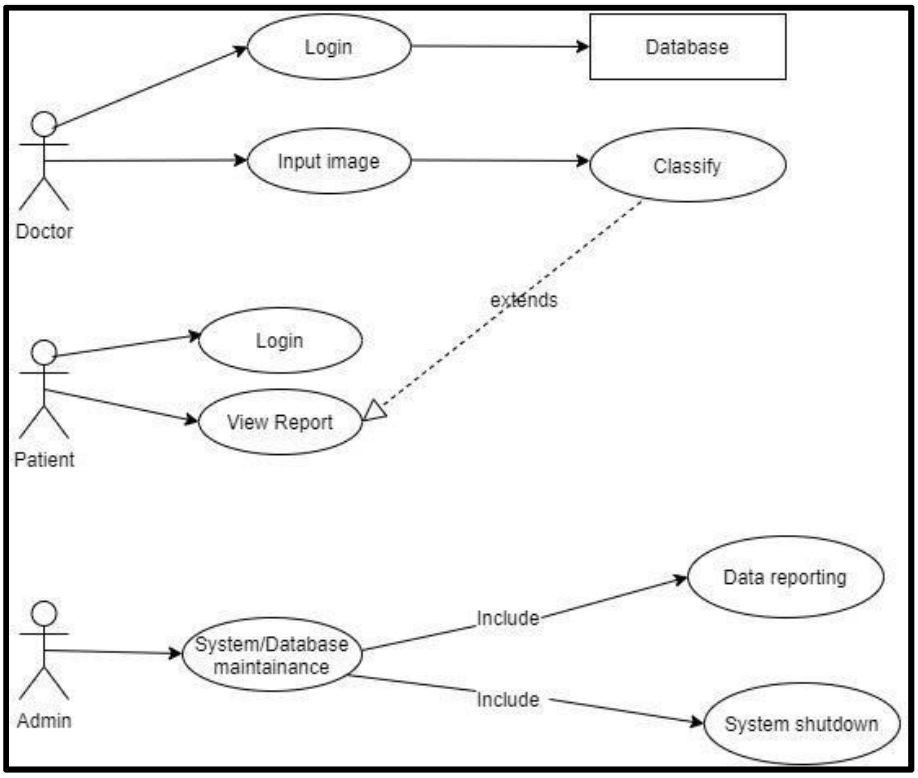

Fig 3: Use Case Diagram

3.2.3 ER-Diagram

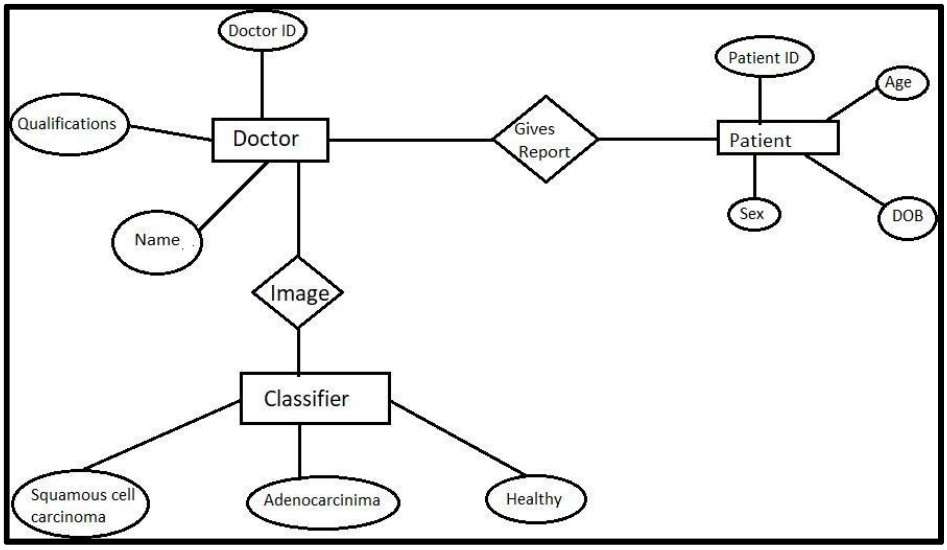

Fig 4: ER-Diagram

3.2.3 Design Template

Registration for Doctor and Medical Practitioner to work on the product

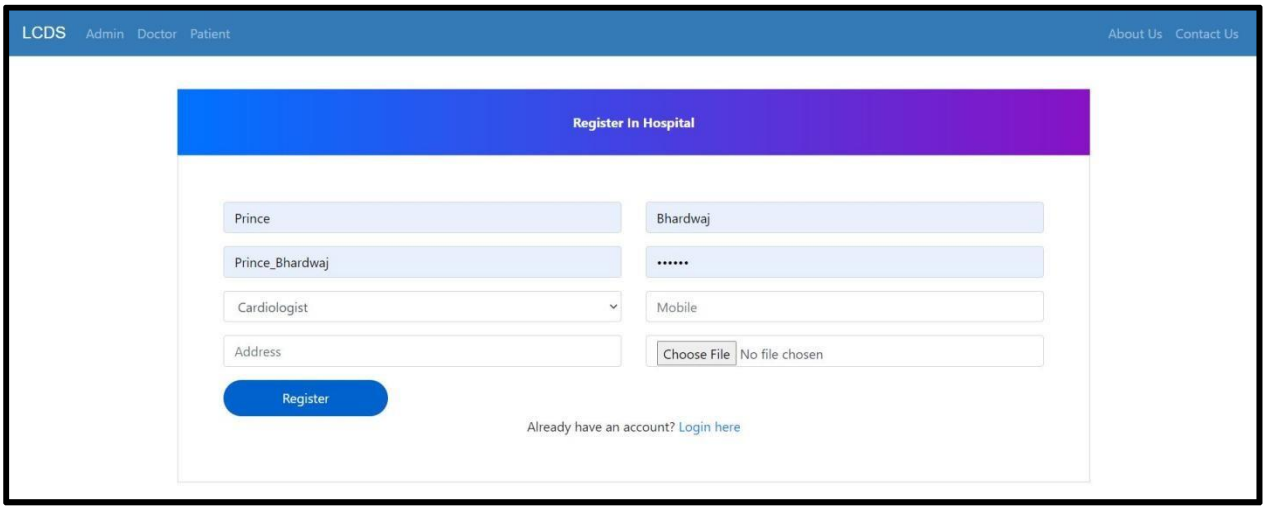

Fig 5: Registration form

After registration It will lead to a page where the doctor and medical practitioner have to provide their credentials to work on the Product Web App

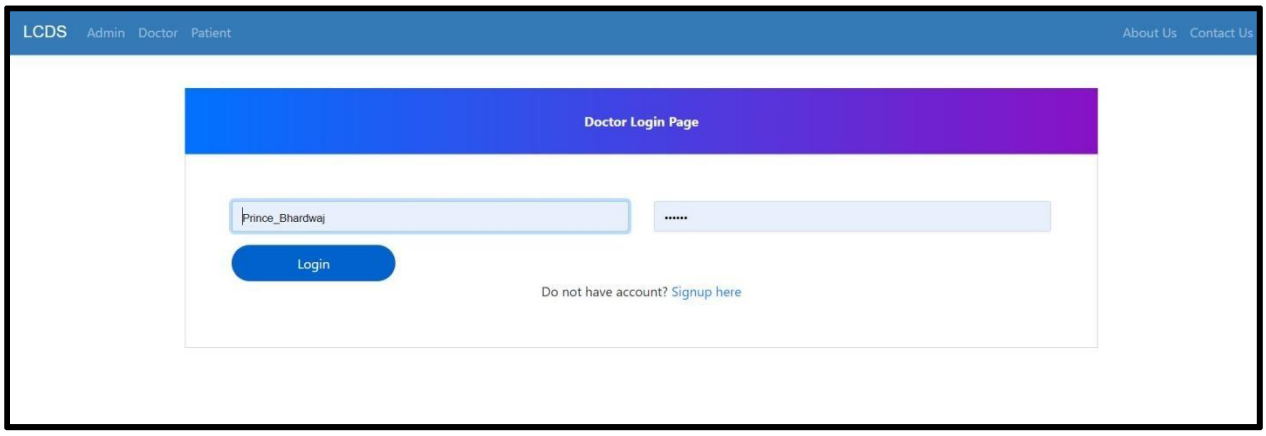

Fig 6: Login Form

After uploading Colored image, Xray etc which shows Adenocarcinomas Cancer. The image is taken as an input and is loaded into the program and will Classify Between the type of Lung Cancer and will give output as ACA Cancer Category with classification report.

3.3 Design and Test Steps/Criteria

3.3.1 Process Model

A process is grouped into a model of the same kind by the Process Model. As a consequence, a model describes a mechanism on a type-level basis. Even though, paradigm has now reached the type stage, it is still a process of instantiation. The same method model is often used to create multiple iterations and has various instantiations. A system model should be used to prescribe how tasks can be carried out concerning the currently taking place.

The objective of a model is as follows:

• Descriptive:

1. Keep track of what occurs during a procedure.

2. Consider an outside expert's perspective who examines how an operation is carried out and determines whether changes are to be made to make it more successful or reliable.

• Prescriptive:

1. Definition of the procedures needed including how they're being performed.

2. Set laws, procedures and patterns of action that will contribute to the desired performance of the process if applied. It can vary between strict adherence and fluid guidance.

• Explanatory:

1. Provide details on the rationale behind such methods.

2. Centered on logical reasoning, analyse and compare various potential courses of action.

3. Make a strong connection between both the procedures and the standards which that model would meet.

4. Also before the the positions at which tracking data can be obtained.

3.3.1.1 Incremental Model

The loop model in our method is seen as an incremental solution. (Pictured) On the basis of the design and implementation of the project is chosen the S/w engineering process model. We have chosen an Incremental Model for our project.

A small collection of specifications are enforced easily and distributed to the authority/customer using the Incremental model.

• Changed & extended demands can be added step by step.

• It combines elements with the iterative prototyping theory of the linear sequential paradigm.

• A deliverable increment of the S/w is generated by each linear sequence.

• The Linear Sequence is divided into four sections:-

1. **Analysis:** Device & software specifications are reported and reviewed.

2. **Design:** Includes four software attributes: Data structure, S/w Architecture, representation of the interface & procedural information.

3. **Coding:** This step is used to convert the design into machine code.

4. **Testing:** Works with S/w logical internals and guarantees that all declarations are right to detect all secret errors.

Advantages of Incremental Model:

• Generates S/w function rapidly & early during the life cycle.

• More versatile & less expensive for changing specifications.

• Easier for checking & debugging

• Customers will react to each designed product.

Why is the Incremental Approach used?

In order to boost the project's performance and usability, the key aim of using the model is to add additional features to the current modules. Using this model, we will adapt to changing consumer needs, which helps to expand the project in a very short period. The next increment in the previous raise incorporates input from consumers and several extra requirements. The process is replicated before the project is completed.

Characteristics of Incremental Model:

1. These models allow the rapid implementation and delivery of a new set of industry requirements to clients and then updating and expanding functionality step by step.

2. Each increment generates the commodity sent to the consumer and proposes certain adjustments and increments that differ with certain extra criteria compared to previous ones.

3. The radical model prevents the initiative from being completed all at once. This is useful for designing and checking components, enabling the project to be modularized for easier management. Ultimately, the growth of the project in increments is easier. We will create a working prototype form 1 with only core tasks and then in subsequent increments, expand on this

layout. By splitting the entire system into separate priority groups, this will serve to reduce system complexity.

Chapter 4: DEVELOPMENT AND IMPLEMENTATION

The aim of this project is to develop a system which is able to classify non-small cell lung cancer into squamous cell carcinoma and adenocarcinoma. To develop this system we used deep learning for developing a model which is able to classify between lung cancer types: Squamous cell carcinoma and Adenocarcinoma. And for front-end we used Django.

4.1 Algorithm used:-

A three-layer CNN algorithm is used to develop this model. Convolutional Neural Networks methods are used because they have good capability to extract features in images without complex pre-processing.

Layers of CNN - First layer is the convolutional layer second is the pooling layer and the third layer is a fully connected layer that comprises a three-layer CNN.

Convolutional layer - CNN's main building element is the convolution layer. It carries the majority of the computational cost on the network.

Pooling layer - By computing a summary statistic of neighbouring outputs, the pooling layer replaces network output at certain spots. In turn, this reduces the size of the representation, which reduces the computation and weights required.

Fully connected layer - As in a conventional fully convolution neural network, the neurons in the fully connected layer are completely linked to all neurons in the preceding and following layers. The fully connected layer assists in the representation mapping between input and output.

CNN architecture is utilized and it is shown in figure no. In this methodology, there are three layers, each with its own specification, and each with a different filter size for feature extraction, Max Pooling (for preserving features), Flattening, and kernels of 7x7x7 dimensions. This model is obtained from a standard CNN by replacing the fully connected layers with a set of 1x1x1 convolutional filters.

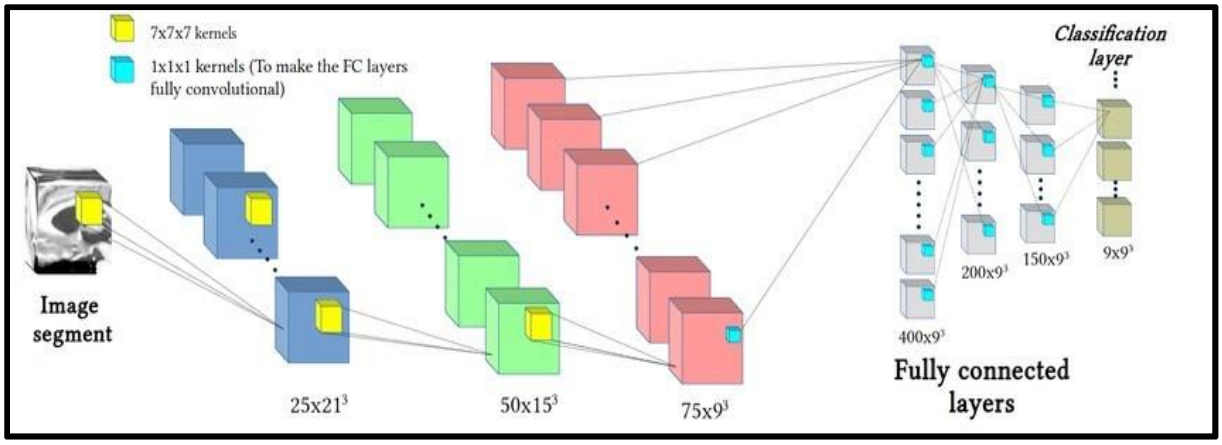

Fig 7: CNN architecture

4.2 Activation functions:-

An activation function calculates the weighted sum of inputs and biases of a neural network, which determines whether a neuron can be activated or not.

The activation function that are used in our model are Relu and Softmax.

```
model = Sequential()
model.add(Conv2D(32, (5, 5), input_shape=input_shape, activation='relu'))
model.add(BatchNormalization())
model.add(MaxPooling2D(pool_size=(3, 3)))

model.add(Conv2D(64, (3, 3),activation='relu'))
model.add(BatchNormalization())
model.add(MaxPooling2D(pool_size=(2, 2)))

model.add(Conv2D(128, (3, 3),activation='relu'))
model.add(BatchNormalization())
model.add(MaxPooling2D(pool_size=(2, 2)))

model.add(Flatten())
model.add(Dense(256,activation='relu'))
model.add(BatchNormalization())
model.add(Dropout(0.25))
model.add(Dense(128,activation='relu'))
model.add(BatchNormalization())
model.add(Dense(num_classes,activation='softmax'))
model.summary()
```

Fig 8: Activation functions

ReLU - ReLU stands for rectified linear unit and It is a activation function. Most neural networks, especially CNNs, utilize ReLU as an activation function.

Fig 9: ReLU

Advantage of Relu-

Utilizing the ReLU function over other activation functions has the main advantage of

not activating all the neurons at once.

Softmax- There is a mathematical function called Softmax, that converts a vector of numbers into a vector of probabilities.

Advantage of using Softmax-

Softmax's main advantage is the range of output probabilities. The range will be 0 to 1, and the sum of all probabilities will equal one. When the softmax function is applied to the multi-classification model, it returns probabilities for each class, and the target class will have a high probability.

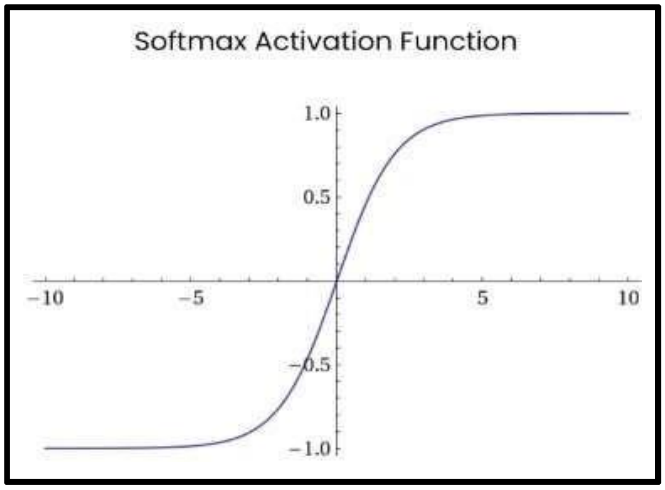

Fig 10: softmax

4.3 Dataset:-

The LC25000 dataset is utilised, which may be acquired from Kaggle. There are 25,000 colour photos in this dataset, divided into five groups of 5,000 images each. The image resolution of all the images are 768 by 768 pixels which are in jpeg file format. Our dataset is available as a 1.85 GB zip file called LC25000.zip. [10] The primary folder lung colon image set comprises two

subfolders: colon image sets and lung image sets, which can be found after unzipping. The colon image sets subdirectory has two secondary subfolders: colon aca, which has 5,000 photos of colon adenocarcinomas, and colon n, which contains 5,000 images of benign colonic tissues. The lung image sets subfolder contains three secondary subfolders: lung aca, lung scc, and lung n. The lung aca subfolder contains 5,000 photographs of lung adenocarcinomas, lung scc contains 5,000 images of lung squamous cell carcinomas, and lung n contains 5,000 images of benign lung tissues.

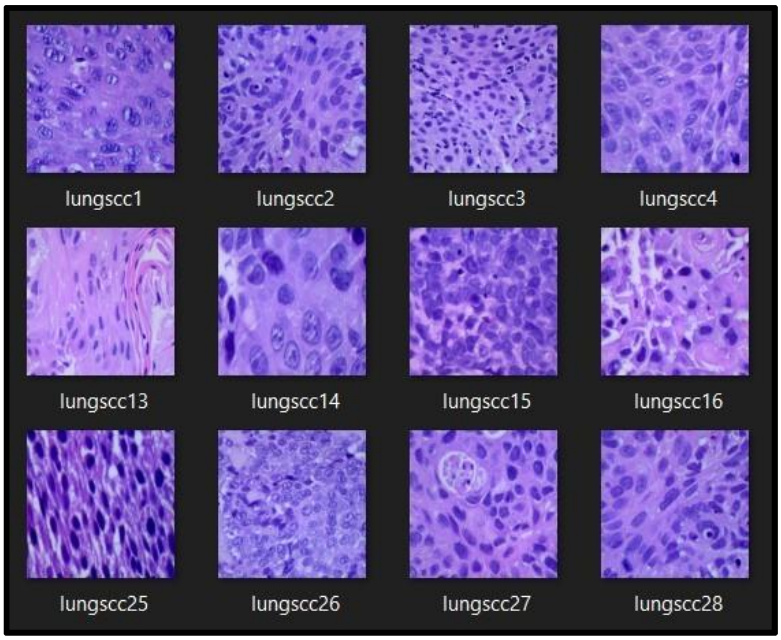

Fig 11: Squanomous Cell Carcinoma image samples

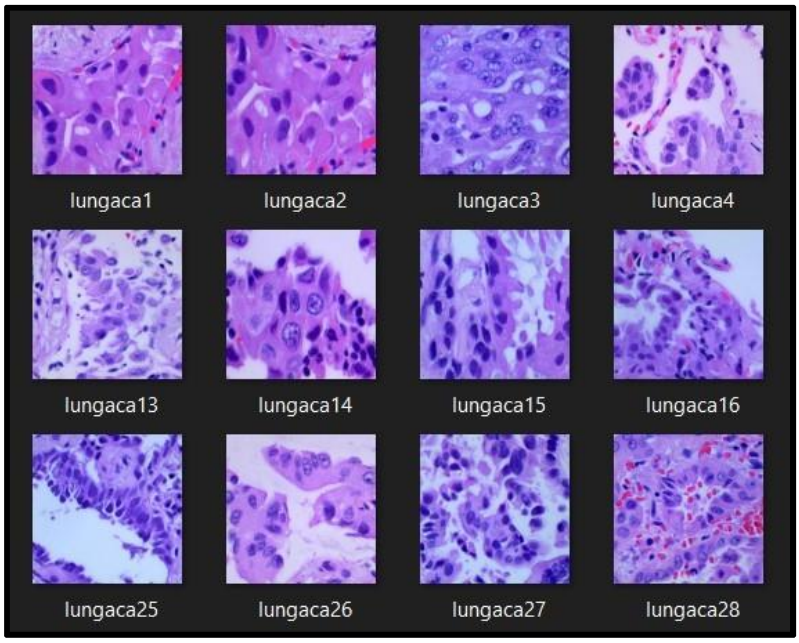

Fig 12: Adenocarcinoma image samples

4.4 Libraries used:-

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import glob
import os

import tensorflow as tf
import keras
from keras.models import Model
from keras.layers import Dense,Dropout,Flatten

from keras.preprocessing.image import ImageDataGenerator
from keras.layers import Conv2D,MaxPooling2D,Activation,AveragePooling2D,BatchNormalization
from keras.models import Sequential

import warnings
warnings.filterwarnings("ignore")
print("Module and Packages imported successfully.....!")
```

Fig 13: Libraries Used

Numpy, pandas, matplotlib, glob and os libraries are used.

Numpy:-

NumPy makes complex numerical operations easy using an open-source Python library. Complex numerical operations must be performed with large datasets when working with machine learning and deep learning applications. When compared to their pure Python implementation, NumPy makes it relatively easy and effective to implement these operations.

Pandas:-

Pandas is an open-source python package built on top of Numpy. It is used as one of the most important data cleaning and analysis tool. It is associated in manipulation of tabular data in dataframes.

Matplotlib:-

In machine learning, among the most popular and oldest Python plotting libraries the Matplotlib is very popular. It helps to understand the huge amount of data through different visualizations of Machine Learning.

Glob:-

A glob is a python standard library. It is used to return all the paths that matches a specific pattern. Glob returns the list of files with their full path.

Os:-

Os is a python module that is used to provide the functionality for interacting with the operating system.

4.5 Optimizer Used:-

```
opt=tf.keras.optimizers.Adam(lr=0.01)
model.compile(optimizer=opt,loss='categorical_crossentropy',metrics=['accuracy','Recall','Precision','AUC'])
train=model.fit(train_generator,
                          epochs=15,
                          steps_per_epoch=train_generator.samples // batch_size,
                          validation_data=validation_generator,
                          validation_steps= validation_generator.samples // batch_size,
                          verbose=1)
```

Fig 14: Adam optimizer

Adam's optimizer is employed. The Adam Optimizer is a stochastic down gradient variant. The term "adaptive time estimation" is abbreviated as Adam. For gradient-based convex optimization algorithms, it is currently the better solution. Hyperparameters have the disadvantage of being easier to tune.

4.6 Data Preprocessing:-

Data preprocessing :-

```
train_datagen=ImageDataGenerator(rescale=1./255,
                                 shear_range=0.2,
                                 zoom_range=0.2,
                                 validation_split=0.2,
                                 horizontal_flip=True)

validation_datagen=ImageDataGenerator(rescale=1./255,
                                 shear_range=0.2,
                                 zoom_range=0.2,
                                 validation_split=0.2,
                                 horizontal_flip=True)
```

```
img_width,img_height =299,299
input_shape=(img_width,img_height,3)
batch_size =32

train_generator =train_datagen.flow_from_directory(train_dir,
                                                   target_size=(img_width,img_height),
                                                   batch_size=batch_size)
validation_generator = validation_datagen.flow_from_directory(
                                                    val_dir,
                                                    target_size=(img_height, img_width),
                                                    batch_size=batch_size)
```

Fig 15: Data Preprocessing

In image pre-processing, the images in our dataset are cropped to the size of 299x299 pixel from 768x768 pixel to reduce the background area and 299x299 size is suitable for CNN method.
Then the cropped image is used for training and validation dataset on 3 layer CNN. During image pre-processing author set the value of rescale is 1./255, shear_range is 0.2, zoom_range is 0.2 and validation_split is 0.2.

4.7 Epoches used:-

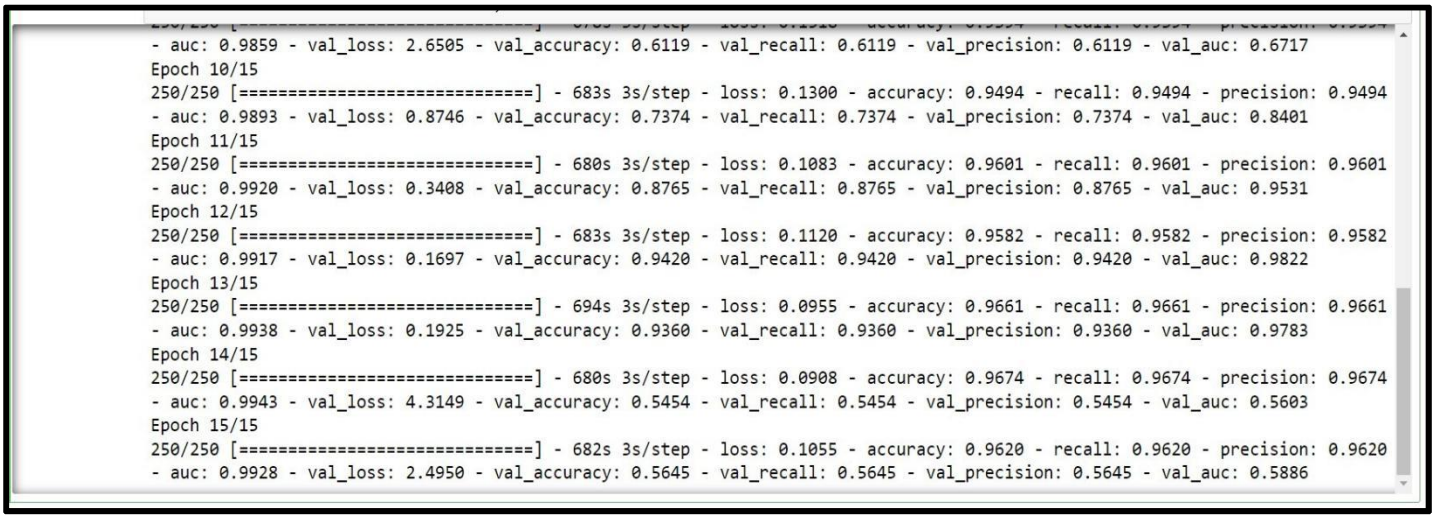

Fig 16: Epochs Used

We used total 15 epoches to get the best accuracy of our model. First we tried with 10 epoches but we can not get as much accuracy as we expect so we decided to go with 15 epoches. After 15 the proposed model observed 96.20% accuracy, 96.20% recall, 96.89% precision, 96.20% AUC and 56.45% validation accuracy over 15 epochs.

4.8 Modules

4.8.1 GUI developed:-

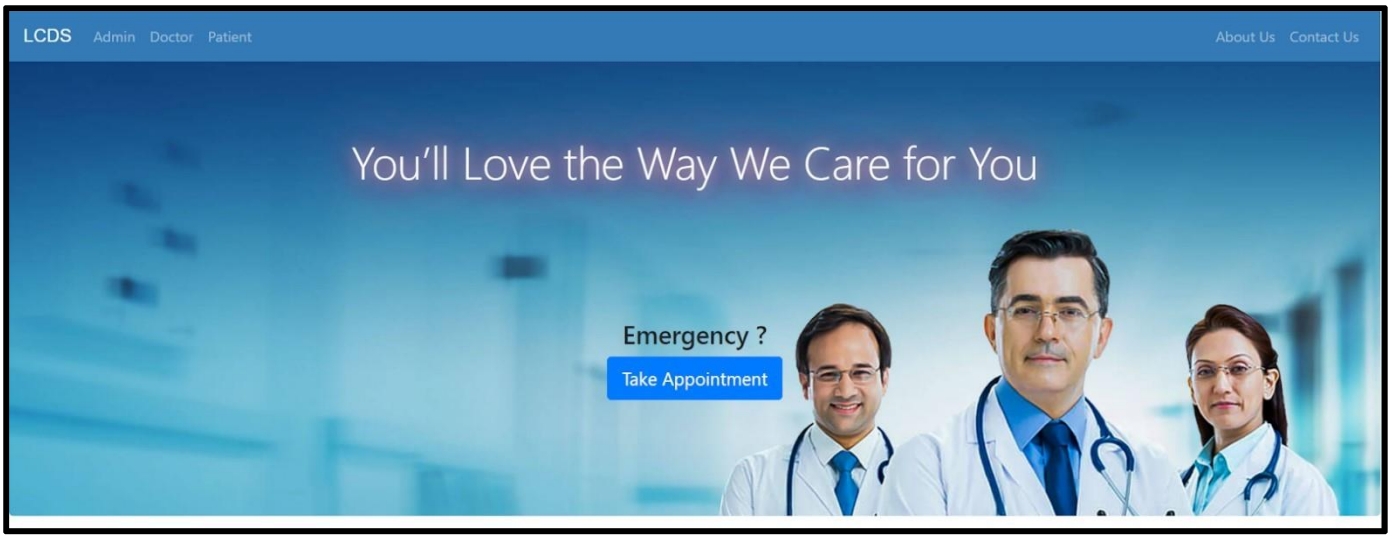

Fig 17: GUI Developed

This is the homepage of our web based application "Lung cancer detection system". In homepage many options will appear like Admin, Doctor, Patients, About Us, Contact Us. Here the doctors and patients can register themselves into the hospital account. About For registering the doctor have to fill their personal details like- Name, Age, Sex, Doctor ID, Qualifications, Address etc. And patients have to fill their details like- Name, Age, Sex, Patient ID, Date of birth, disease(if any) etc. About Us and Contact Us option can give the information about the hospital and how can a patient contact to the hospital by dialing hospital phone numbers or using hospital mail id.

When we scroll down by the homepage then we get another page to login by the Admin, Doctors and Patients.

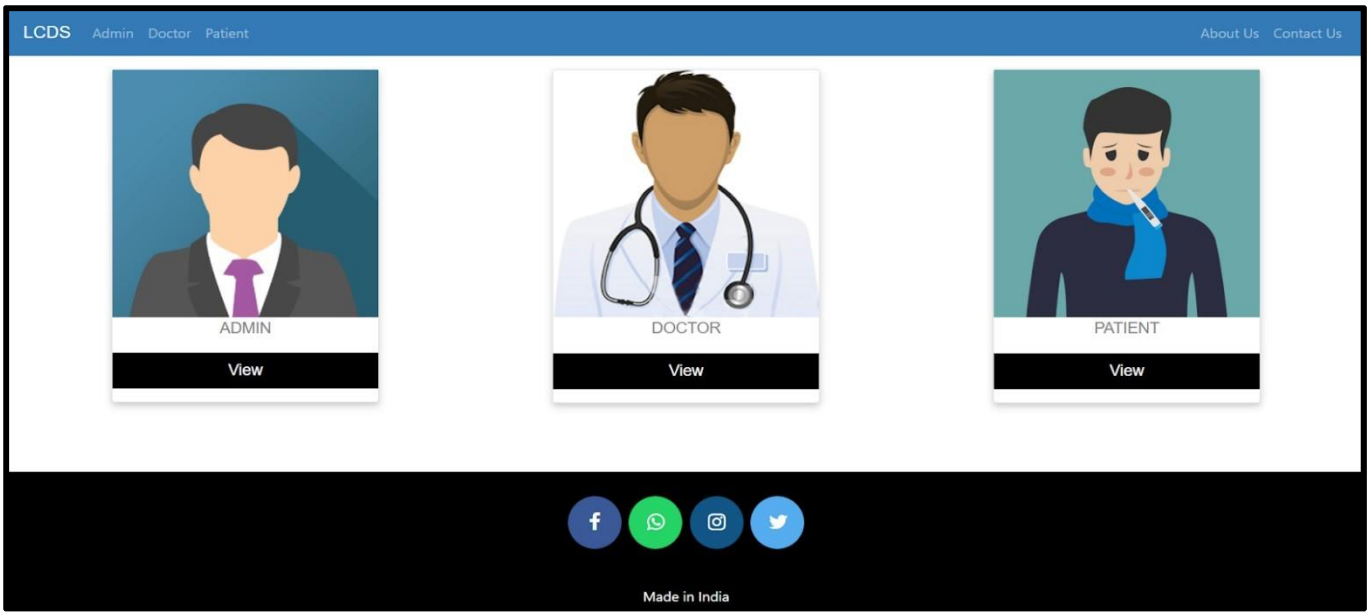

Fig 18: Login Screen

In this page the Admin, Doctor and Patient can login into the website and access the information related to each other.

4.8.2 Admin login:-

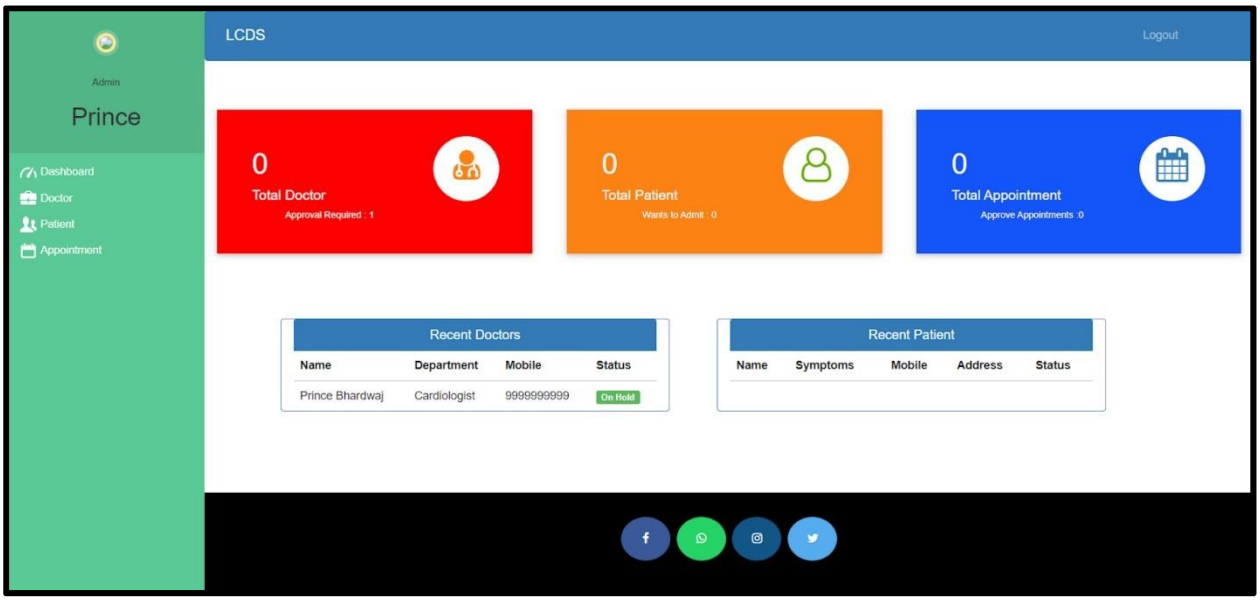

Fig 19: Admin Login

In Admin login the admin can access many i formation of the hospital like-

- How many Doctors are logged into the hospital website.
- How many Patients are logged into the hospital website.
- How many total appointments are generated by the patients to the specific doctor.
- How many Doctors recently login into the website with their name, department, mobile no. and status of the doctor.
- How many Patient recently login into the website with their name, symptoms, mobile no., address and status of the patient.

Functions:-

Admin:-

- Admin can sign-up their account and login.
- Admin can register the doctors.
- Admin can approve, reject or delete the doctors.
- Admin can view the patients.
- Admin can approve and then admit the patient.
- Admin can reject and discharge patient.
- Admin can Generate Invoice in the form of pdf and patient can download it after logging in.
- Admin can approve patients appointments.

Doctor:-

- Doctors can sign-up and then login after getting approval from the admin.
- Doctors can view the patient details like what is the name of the patient, what are the symptoms of the patient.
- Doctor can see the list of the discharged patients.
- Doctor can see the appointments list made by the admin.
- Once Doctor attend the appointment of the patient then they can delete the appointment.

4.8.3 Patient:-

- Patients can sign-up and create account to admit in hospital.
- Patient can login after getting approval from hospital's admin.
- Assigned doctor's details can be assessed by the patients like specialization and mobile number of the doctor.
- Appointment status can be checked by the patients like it is pending or confirmed.
- Appointments of the doctors can be booked by the patients.
- When the patient is discharged then they can access the Invoice after downloading it.

Chapter 5: Testing and Results

5.1 Testing Process

5.1.1 Software Testing

The role of software testing is to ensure that programs are efficient and accurate. Software testing is an observational science investigation conducted to provide consumers with information regarding a product's quality in the environment in which it is intended to function. This can include but is not limited to running a program or application to detect errors.

5.1.2 Unit Testing:-

In this case, each module is evaluated independently. The standards for defining unit test modules were selected to identify modules that have key functionality. A module may be either an individual or a method.

The unit testing functions that will be tested are as follows:

- Provide a movie name.
- Convert the name string to vector.
- Use the vector to generate the similarity score.
- Sort the movies according to a descending similarity score.
- Pick top 10 movies.

5.1.3 Integration Testing:-

Relevant components are integrated and analyzed as a group during integration planning. Integration testing takes unit-tested elements like data, groups them into larger aggregates,

applies integration test plan tests to those aggregates, and produces the integrated testing framework.

5.1.4 Validation Testing:-

At the start or end of the production process, this approach is used to determine if the software satisfies the specified specifications. In the validation testing, all the use cases of products are tested and managed.

5.1.5 GUI Testing:-

GUI testing is the process of examining a product's graphical user interface to ensure that it complies with standards, such as retaining navigation between icons/buttons with source code.

5.2 Results:-

In the Pre-implementation the experiments utilized LC25000 database downloaded from Kaggle and it consist of 25,000 images and the author used 5000 color images of lung squamous cell carcinoma and 5000 color images of lung adenocarcinoma and split ratio between training samples and validation samples is 80 and 20. For binary classification, the dataset consists of 8,000 images from training and 2,000 images from validation. In the data preprocessing authors cropped the images in 299x299 pixels for 3 layers CNN implementation.

The proposed approach uses 3 layers of CNN which is a more efficient neural network architecture to classify the lung cancer images of LC25000 dataset. The proposed model observed 96.89% accuracy, 96.89% recall, 96.89% precision, 99.44% AUC and 93.20% validation accuracy over 15 epochs. Post-implementation profuse and ideal results are obtained as shown in figures 20 and 21.

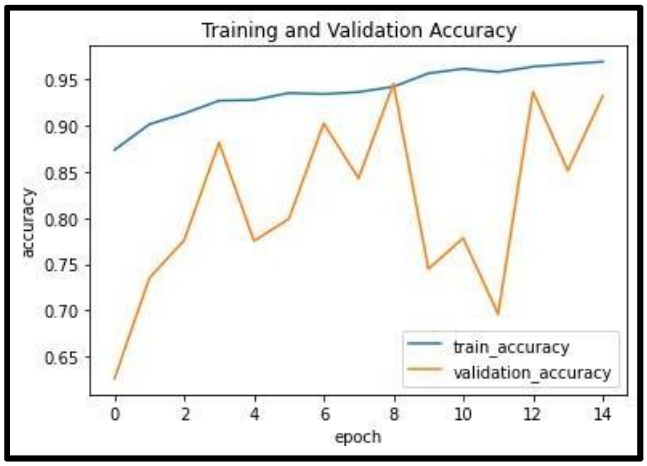

Fig 20: Training and validation accuracy

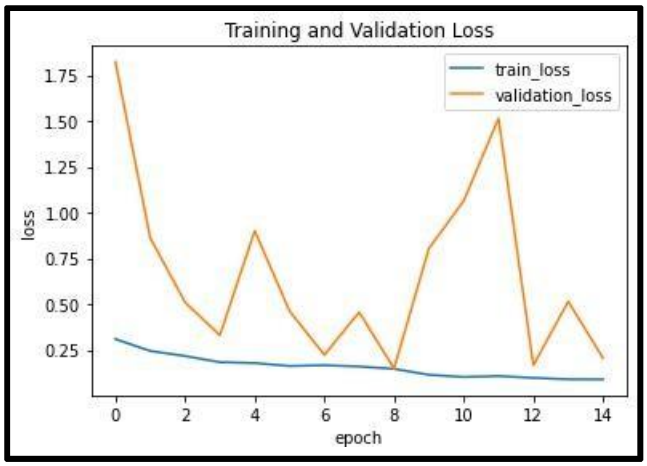

Fig 21: Training and validation loss

The accuracy shown in the graphs presents the encompassing evidence acquired and a satisfactory view of action providing an edge to automate the lung cancer detection and figures 22 and 23 show the confusion matrix of our proposed model.

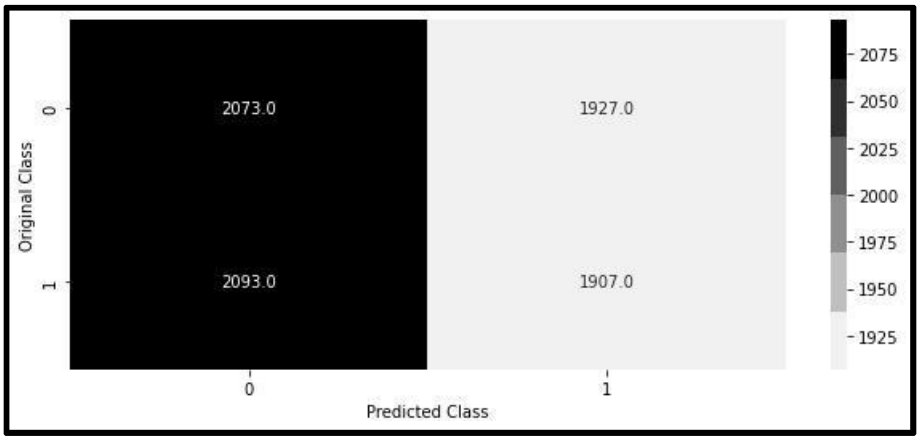

Fig 22: Training confusion matrix

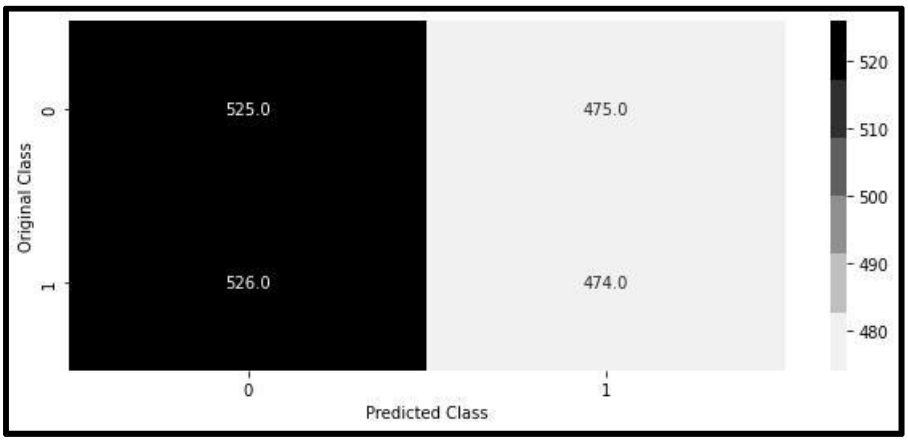

Fig 23: Validation confusion matrix

Figure 24 shows the prediction of the lung cancer type present in the sample image. Here Squamous cell carcinoma is predicted.

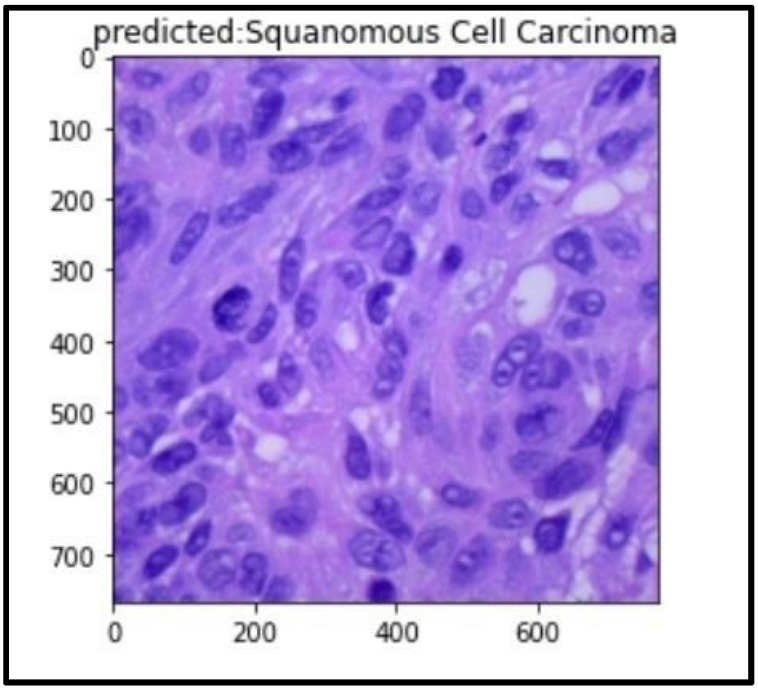

Fig 24: Squamous cell carcinoma prediction

Chapter 6: CONCLUSION

With the proposed CNN, the experiments show the excellent performance with the image dataset containing 5000 color images of lung squamous cell carcinoma and 5000 color images of lung adenocarcinoma from the LC25000 dataset to observed a training accuracy of 96.89% and 93.20% validation accuracy. This research work presents sub-classifying lung cancer using CNN. In this work, the training of the model on previously unseen non-small cell lung cancer images also allowed to classify them into squamous cell carcinoma and adenocarcinoma.

6.1 Further Improvements:-

- Can be used to detect multiple types of lung cancer
- More Datasets can be used to generate the model.
- Use more data features for better performance.
- Use of different dataset for different diseases.

REFERENCES

[1] Makaju S, Prasad PW, Alsadoon A, Singh AK, Elchouemi A. Lung cancer detection using CT scan images. Procedia Computer Science. 2018 Jan 1;125:107-14.

[2] Asuntha A, Srinivasan A. Deep learning for lung Cancer detection and classification. Multimedia Tools and Applications. 2020 Mar;79(11):7731-62.

[3] Teramoto, A., Tsukamoto, T., Kiriyama, Y., & Fujita, H. (2017). Automated classification of lung cancer types from cytological images using deep convolutional neural networks. BioMed Research International, 2017, (pp. 1–6).

[4] Hatuwal BK, Thapa HC. Lung Cancer Detection Using Convolutional Neural Network on Histopathological Images. Int. J. Comput. Trends Technol. 2020;68:21-4.

[5] Tekade R, Rajeswari K. Lung cancer detection and classification using deep learning. In2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA) 2018 Aug 16 (pp. 1-5). IEEE.

[6] Khan A. Identification of Lung Cancer Using Convolutional Neural Networks Based Classification. Turkish Journal of Computer and Mathematics Education (TURCOMAT). 2021 Apr 28;12(10):192-203.

[7] Nasrullah N, Sang J, Alam MS, Mateen M, Cai B, Hu H. Automated lung nodule detection and classification using deep learning combined with multiple strategies. Sensors. 2019 Jan;19(17):3722.

[8] Asuntha A, Srinivasan A. Deep learning for lung Cancer detection and classification. Multimedia Tools and Applications. 2020 Mar;79(11):7731-62.

[9] Alakwaa W, Nassef M, Badr A. Lung cancer detection and classification with 3D convolutional neural network (3D-CNN). Lung Cancer. 2017;8(8):409.

[10] Masood A, Sheng B, Yang P, Li P, Li H, Kim J, Feng DD. Automated decision support system for lung cancer detection and classification via enhanced RFCN with multilayer fusion RPN. IEEE Transactions on Industrial Informatics. 2020; 21;16(12):7791-801.

[11] Perumal S, Velmurugan T. Lung cancer detection and classification on CT scan images using enhanced artificial bee colony optimization. International Journal of Engineering & Technology. 2018;7(2.26):74-9.

[12] Zhang C, Sun X, Dang K, Li K, Guo XW, Chang J, Yu ZQ, Huang FY, Wu YS, Liang Z, Liu ZY. Toward an expert level of lung cancer detection and classification using a deep

convolutional neural network. The oncologist. 2019 Sep;24(9):1159.

[13] S. Hawkins, H. Wang, Y. Liu, A. Garcia, O. Stringfield, H. Krewer, Q. Li, D. Cherezov, R. A. Gatenby, Y. Balagurunathan, D. Goldgof, M. B. Schabath, L. Hall, and R. J. Gillies, "Predicting malignant nodules from screening ct scans," Journal of Thoracic Oncology, vol. 11, no. 12, pp. 2120–2128, 2016.

[14] Y. Bar, I. Diamant, L. Wolf, S. Lieberman, E. Konen, and H. Greenspan, "Chest pathology detection using deep learning with non-medical training," Proceedings - International Symposium on Biomedical Imaging, vol. 2015-July, pp. 294–297, 2015.

[15] J. Tan, Y. Huo, Z. Liang, and L. Li, "A comparison study on the effect of false positive reduction in deep learning based detection for juxtapleural lung nodules: CNN vs DNN," in Proceedings of the Symposium on Modeling and Simulation in Medicine, MSM '17, pp. 8:1–8:8, Society for Computer Simulation International, 2017.

[16] R. Golan, C. Jacob, and J. Denzinger, "Lung nodule detection in ct images using deep convolutional neural networks," in 2016 International Joint Conference on Neural Networks, pp. 243–250.

[17] D. Kumar, A. Wong, and D. A. Clausi, "Lung nodule classification using deep features in CT images," in 2015 12th Conference on Computer and Robot Vision, pp. 133–138, June 2015.

[18] A. Chon, N. Balachandar, and P. Lu, "Deep convolutional neural networks for lung cancer detection," tech. rep., Stanford University, 2017.

[19] W.-J. Choi and T.-S. Choi, "Automated pulmonary nodule detection system in computed tomography images: A hierarchical block classification approach," Entropy, vol. 15, no. 2, pp. 507–523, 2013.

[20] H. Han, L. Li, H. Wang, H. Zhang, W. Moore, and Z. Liang, "A novel computer-aided detection system for pulmonary nodule identification in CT images," in Proceedings of SPIE Medical Imaging Conference, vol. 9035, 2014.