

2023

UG Project Report on

**PATHOLOGY IMAGE ANALYSIS FOR LUNG CANCER CLASSIFICATION USING IBM WATSON MACHINE LEARNING**



**SUBMITTED BY**

**Srihitha ch 19UK1A0582**

**PRASHANTH d 19UK1A0581 SHIVA RAMA KRISHNA A 19UK1A0572**

**ANUHYA A 19UK1A05B8**



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**VAAGDEVI ENGINEERING COLLEGE**

(Affiliated to JNTUH, Hyderabad)

Bollikunta, Warangal – 506005

**2019– 2023**

 | [Company address]

**CONTENTS**

**CHAPTERS PAGE NO’S**

1. INTRODUCTION……………………………………………………….02-03
   1. OVERVIEW……………………………………………………………..02
   2. PURPOSE……………………………………………………………….03
2. LITERATURE SURVEY…………………………………………………..04
   1. EXISTING PROBLEM…………………………………………………04
   2. PROPOSED SOLUTION………………………………………………..04

**3. THEORETICAL ANALYSIS 05-06**

3.1 PROJECT ARCHITECTURE… 05

* 1. BLOCK DIAGRAM…………………………………………………….06
  2. HARDWARE AND SOFTWARE REQUIREMENTS 06

4. [EXPERIMENTAL INVESTIGATIONS 07](#_TOC_250003)

5. FLOWCHART AND ARCHITECHTURE…………………………...…08-09

6. RESULT…………………………………………………………………....11-13

7. ADVANTAGES AND DISADVANTAGES……………………………..….14

8. APPLICATIONS…………………………………………………………..…15

9. CONCLUSION AND FUTURE SCOPE…………………………………....16

10. BIBILIOGRAPHY………………………………………………………….17

**CHAPTER 1**

**INTRODUCTION**

* 1. **OVERVIEW**

 Automatic defects detection in CT images is very important in many diagnostic and therapeutic applications. Because of high quantity data in CT images and blurred boundaries, tumor segmentation and classification is very hard. This work has introduced one automatic lung cancer detection method to increase the accuracy and yield and decrease the diagnosis time. The goal is classifying the tissues to three classes of normal, benign and malignant. In MR images, the amount of data is too much for manual interpretation and analysis. During past few years, lung cancer detection in CT has become an emergent research area in the field of medical imaging system. Accurate detection of size and location of lung cancer plays a vital role in the diagnosis of lung cancer. The diagnosis method consists of four stages, pre-processing of CT images, feature, extraction, and classification, the features are extracted based on DTCWT and PNN. In the last stage, PNN employed to classify the Normal and abnormal.

**1.2 PURPOSE**

Cancer is a group of diseases characterized by the uncontrolled growth and spread of abnormal cells. If the spread is not controlled, it can result in death. Lung cancer was the most common cancer in worldwide, contributing 2,093,876 of the total number of new cases diagnosed in 2018.

The incidence rate has been declining since the mid- 1980s in men, but only since the mid-2000s in women, because of gender differences in historical patterns of smoking uptake and cessation.

From 2005 to 2015, lung cancer incidence rates decreased by 2.5% per year in men and 1.2% per year in women. Symptoms do not usually occur until the cancer is advanced, and may include persistent cough, sputum streaked with blood, chest pain, voice change, worsening shortness of breath, and recurrent pneumonia or bronchitis.

Cigarette smoking is by far the most important risk factor for lung cancer; 80% of lung cancer deaths in the US are still caused by smoking. Risk increases with both quantity and duration of smoking. Cigar and pipe smoking also increase risk. Exposure to radon gas released from soil and building materials is thought to be the second-leading cause of lung cancer in the US. Other risk factors include occupational or environmental exposure to secondhand smoke, asbestos (particularly among smokers), certain metals (chromium, cadmium, arsenic), some organic chemicals, radiation, air pollution, and diesel exhaust. Some specific occupational exposures that increase risk include rubber manufacturing, paving, roofing, painting, and chimney sweeping. Risk is also probably increased among people with a history of tuberculosis. Genetic susceptibility (e.g., family history) plays a role in the development of lung cancer, especially in those who develop the disease at a young age.

We can cure lung cancer, only if you identifying the yearly stage. So here, we use machine learning algorithms to detect the lung cancer. This can be made faster and more accurate. In this study we propose machine learning strategies to improve cancer characterization. Inspired by learning from



**CHAPTER 2**

**LITERATURE SURVEY**

**2.1 EXISTING PROBLEM**

**To complete this project, you should have the following software’s and packages**

**Anaconda Navigator:**

Anaconda Navigator is a free and open-source distribution of the Python and R programming languages for data science and machine learning related applications. It can be installed on Windows, Linux, and macOS. Conda is an open-source, cross-platform, package management system. Anaconda comes with so very nice tools like JupyterLab, Jupyter Notebook, QtConsole, Spyder, Glueviz, Orange, Rstudio, Visual Studio Code. For this project, we will be using Jupiter notebook and spyder.

**2.2 PURPOSED SOLUTION**

The method or solution is  Jupyter notebook and spyder, that are used to complete this project.

**To build Machine learning models you must require the following packages**

**Sklearn:** Scikit-learn is a library in Python that provides many unsupervised and supervised learning algorithms.

**NumPy:** NumPy is a Python package that stands for 'Numerical Python'. It is the core library for scientific computing, which contains a powerful n-dimensional array object

**Pandas:**pandas is a fast, powerful, flexible, and easy to use open source data analysis and manipulation tool,built on top of the Python programming language.

**Matplotlib:**  It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits

**CHAPTER 3**

**THEORETICAL ANALYSIS**

**3.1 BLOCK DIAGRAM**

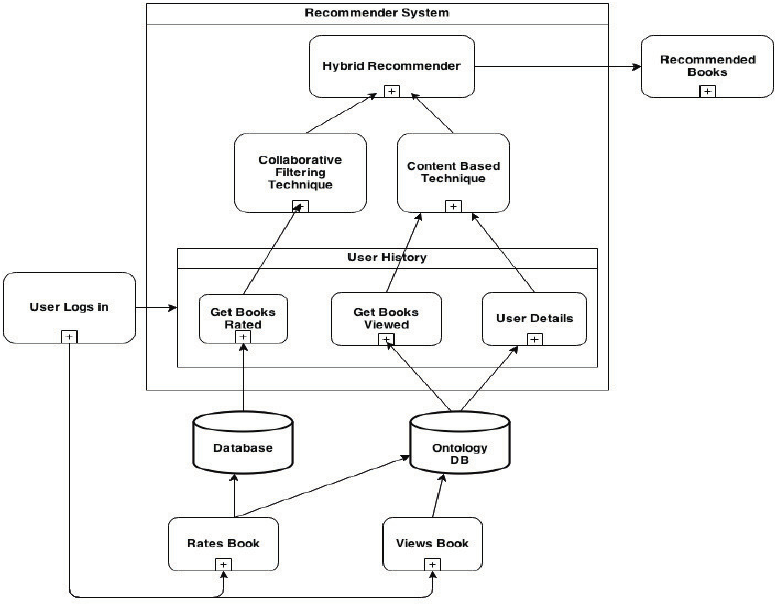
In this  Block Diagram Book Recommendation System Using IBM watson, says recommended by users, for example in Institute or College have college libraries is there, in library have lot of books and different Departments books are stored. the college students have college library ID card must and should, those students are goto college library and what the students want Books they access it come back to class, Some Students don't have college library ID cards.

Fig3.1 Block diagram of Book Recommendation System Using IBM Watson.

because those students are joined college later, those students want books. that situation students are use the college lecture library ID card and access the class Books. this is the recommended of their class teachers to access their class books. In diagram Book recommended of both Users. in this diagram have two Users are there, the User is Accessed the that book and read it, after the Similar User are read the same book, before taking the recommended by first user, because the first user is already read that book. that's why the second user is take the recommendation and read the book, recommendation is how the book is good or bad, and good things or not and what is story

**3.2 HARDWARE / SOFTWARE DESIGNING**

The hardware required for the development of this project is:

      Processor : Intel CoreTM i5-9300H

      Processor speed : 2.4GHz

      RAM Size : 8 GB DDR

      System Type : X64-based processor

**SOFTWARE DESIGNING**:

The software required for the development of this project is:

      Desktop GUI       :  Anaconda Navigator

      Operating system   : Windows 10

      Front end        : HTML, CSS,JAVASCRIPT

      Programming   : PYTHON

      Cloud Computing Service  : IBM Cloud Services

**CHAPTER 4**

**EXPERIMENTAL INVESTIGATION**

**IMPORTING AND READING THE DATASET**

**Importing the Libraries**

First step is usually importing the libraries that will be needed in the program.

**Pandas:** It is a python library mainly used for data manipulation.

**NumPy:** This python library is used for numerical analysis.

**Matplotlib and Seaborn:** Both are the data visualization library used for plotting graph which will help us for understanding the data.

**csr\_matrix() :**A dense matrix stored in a NumPy array can be converted into a sparse matrix using the CSR representation by calling the csr\_matrix() function.

**Train\_test\_split:** used for splitting data arrays into training data and for testing data.

**Pickle:** to serialize your machine learning algorithms and save the serialized format to a file.

**Reading the Dataset**

For this project, we make use of three different datasets (Books\_Ratings, Books, Users). We will be selecting the important features from these datasets that will help us in recommending the best results.

The next step is to read the dataset into a data structure that’s compatible with pandas.  
 Let’s load a .csv data file into pandas. There is a function for it, called **read\_csv().**We will need to locate the directory of the CSV file at first (it’s more efficient to keep the dataset in the same directory as your program).If the dataset in same directory of your program, you can directly read it, without any path. After the next Steps we made following bellow:

1.Data visualization

2.Collaborative and filtering

3.Creating the Model

4.Test and save the model

5.Build Python Code

6.Build HTML Code

7.Run the Application

We are the following above sections we did and investigate it.

**CHAPTER 5**

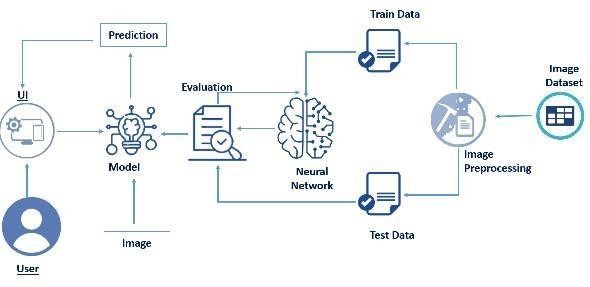
**FLOWCHART AND ARCHITECTURE**

Fig 5.1 Flowchart of the project

**Project Flow:**

* User interacts with the UI (User Interface) to upload the input features.
* Uploaded features/input is analyzed by the model which is integrated
* Once model analyses the uploaded inputs, the prediction is showcased on the UI.

**1. Data Collection.**

* Collect the dataset or Create the dataset

**2. Data Pre- processing.**

* Import the Libraries.
* Importing the dataset.
* Exploratory Data Analysis
* Data Visualization

**3. Collaborating Filtering**

* Merging datasets
* Creating the Model
* Predicting the results
* Saving our model and dataset

**4. Application Building**

* Create an HTML file
* Build a Python Code

**CHAPTER 6**

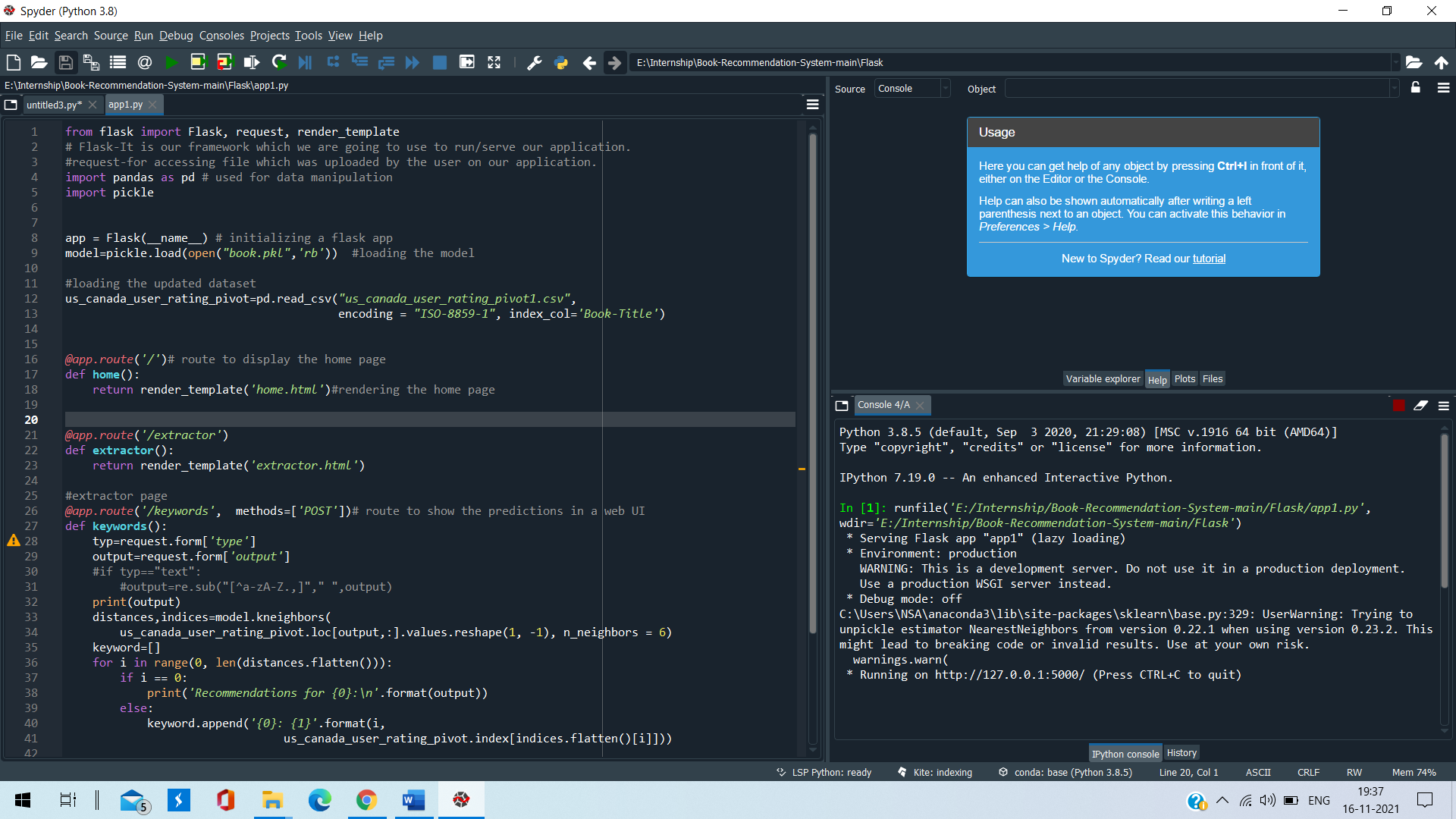
**RESULT**

Fig 6.1 Flask App Code with Output Page

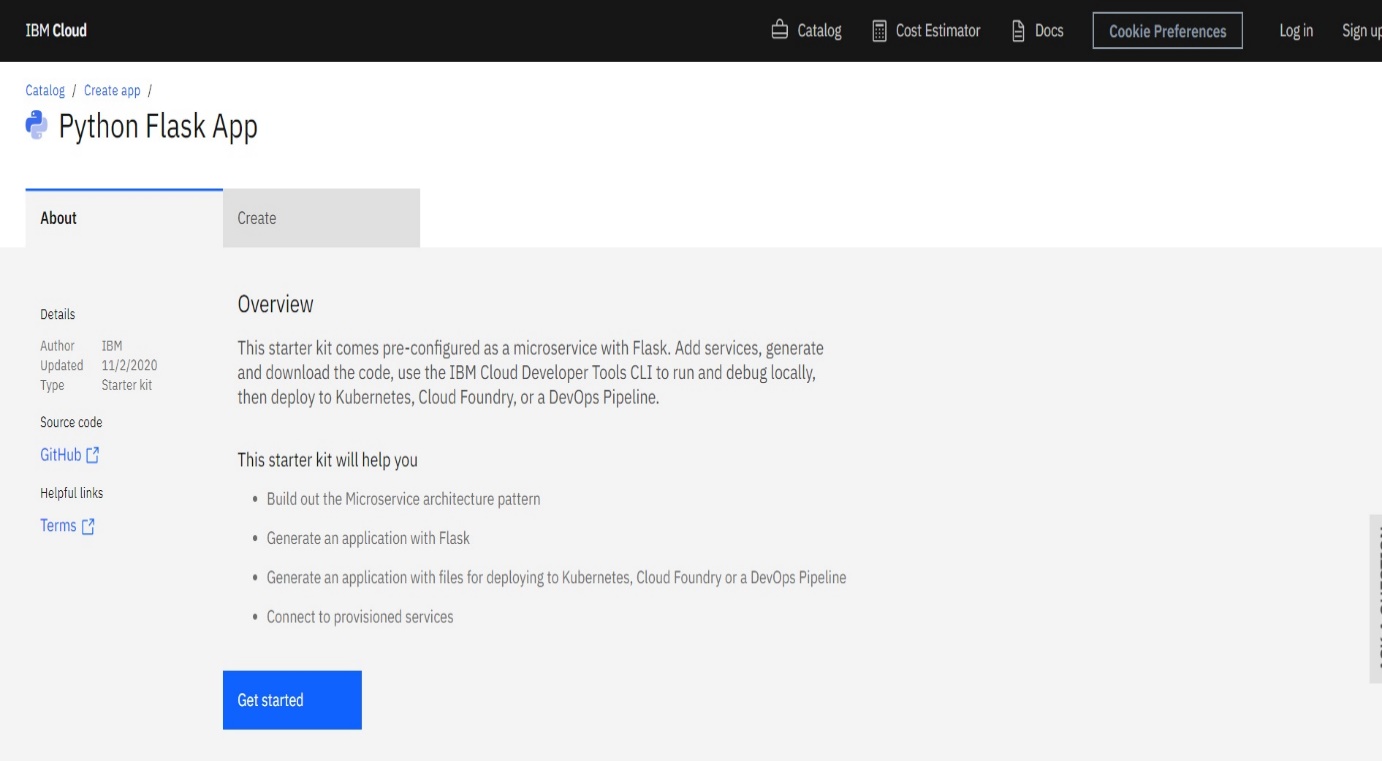


Fig 6.2 Uploading the Project of Book Recommendation System using IBM Watson, In IBM



Fig 6.3 Home Page of Lung Cancer Classification using IBM Watson



Fig 6.4 Extractor Page of Lung Cancer Classification using IBM Watson

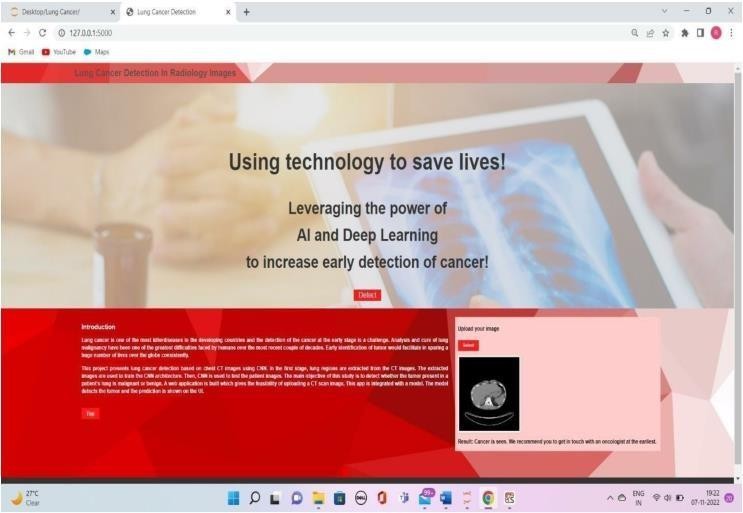


Fig 6.5 Keywords Page of Lung Cancer Classification using IBM Watson

**CHAPTER 7**

**ADVANTAGES AND DISADVANTAGES**

**ADVANTAGES**

High accuracy: Deep learning models have achieved state-of-the-art results in a variety of medical prediction tasks, including lung cancer prediction. They can analyze large amounts of data and identify complex patterns that may not be detectable by humans or other machine learning techniques.

Automation: Deep learning models can automate the prediction process, making it faster and more efficient than manual methods.

Improved patient care: By predicting lung cancer early, doctors can potentially provide treatment earlier, which may improve patient outcomes.

Reduced cost: Automating the prediction process with deep learning can reduce the need for manual labor, potentially lowering the cost of care.

Scalability: Deep learning models can be trained on large datasets and can be easily deployed to make predictions on new data, making them scalable to use in a clinical setting.

**DISADVANTAGES**

Limited interpretability: One of the main limitations of deep learning models is that they can be difficult to interpret. It can be challenging to understand how the model arrived at a particular prediction, which can make it difficult to identify errors or biases in the model.

Dependence on high-quality data: Deep learning models rely on large amounts of high-quality data to learn patterns and make predictions. If the data is not representative or is of poor quality, the model's performance may be compromised.

Limited generalizability: Deep learning models are designed to perform well on the specific task they were trained for. They may not perform as well when applied to different tasks or when applied to data that is significantly different from the training data.

**CHAPTER 8**

**APPLICATIONS**

Deep learning can be used for predicting lung cancer in several ways. One approach is to use deep learning to analyze medical images, such as CT scans, to identify abnormalities that may indicate the presence of lung cancer. This can be done using a convolutional neural network (CNN) trained on a dataset of CT scans with and without lung cancer. The CNN can learn to recognize patterns in the images that are characteristic of lung cancer, such as nodules or masses, and use this information to make a prediction.

Another approach is to use deep learning to analyze other types of data, such as patient medical records, to predict the likelihood of a patient developing lung cancer. This can be done using a machine learning model trained on a dataset of patient records that includes information about risk factors for lung cancer, such as age, smoking history, and family medical history. The model can learn to identify patterns in the data that are associated with an increased risk of lung cancer and use this information to make a prediction.

It is important to note that deep learning models can only make predictions based on the data they are trained on. Therefore, it is important to use a diverse and representative training dataset to ensure that the model is able to generalize to a wide range of patients.

**CHAPTER 9**

**CONCLUSION AND FUTURESCOPE**

**CONCLUSION**

Lung cancer is one of the most dangerous diseases and the most common cause of death, the severity of the disease lies in the difficulty of diagnosing it in the early stages. This paper tries to endeavor to investigate of three classifiers to find the best classifier could classify lung cancer in early stage. The informational indices included in this study were derived from UCI databases for lung cancer patients. The focus of this paper is on using WEKA Tool to investigate the accuracy of classification algorithms. The results show that the Support Vector Machine (SVM) give the best accuracy 95.56%, that can detect lung cancer in its early stages and save several lives and, KNearest Neighbor KNN It gave less accuracy 88.40%.

**FUTURESCOPE**

Several machine learning techniques have been used to detect, predict, compare, or classify lung cancer. However, there are lessons and research directions to advance research in lung and related diseases:

Several models have been used to detect, predict, compare or classify lung cancer using machine learning techniques. However, research is scarce in the use of modern soft computing to provide a high degree of accuracy. Soft computing such as ABC, PSO, Genetic algorithm, functional approximation etc., as a single or hybrid model to understand their strength and weakness in such domain.

Research works into the applications of machine learning in lung cancer are increasing. However, most of the works focused on early detection; there is a need to advance to severity level and other lung cancer components to support medical practitioners in the daily works. Most of the research works used CT scan images to detect, predict, compare or classify lung cancer using classical machine learning techniques, and there is a need to use other data set such as family history, personal attributes and X-ray images to understand if they can provide insights into the presence of lung cancer and related diseases.

**CHAPTER 10**

**BIBILOGRAPHY**

1. Arnaud A. A. Setio, Francesco Ciompi, Geert Litjens, Paul Gerke, Colin Jacobs, Sarah J. van Riel, Mathi”Pulmonary nodule detection in CT images: false positive reduction using multi- view convolutional networks”(2016).
2. Junyuan Xie, Ross Girshick “Unsupervised Deep Embedding for Clustering Analysis”(2016).
3. Mario Buty1, Ziyue Xu1 , Mingchen Gao” Characterization of Lung Nodule Malignancy using Hybrid Shape and Appearance Features”(2017)
4. Alan L. Yuille4 “ Deep Supervision for Pancreatic Cyst Segmentation in Abdominal CTScans”(2018)
5. Kumar, D., Wong, A., Clausi, D.A.: Lung nodule classification using deep features in CT images. In: Computer and Robot Vision (CRV), 2015 12th Conference on. pp. 133–138. IEEE (2015)

**APPENDIX**

**A Source Code of Flask:**

from flask import Flask, request, render\_template

# Flask-It is our framework which we are going to use to run/serve our application.

#request-for accessing file which was uploaded by the user on our application.

import pandas as pd # used for data manipulation

import pickle

app = Flask(\_\_name\_\_) # initializing a flask app

model=pickle.load(open("book.pkl",'rb')) #loading the model

#loading the updated dataset

us\_canada\_user\_rating\_pivot=pd.read\_csv("us\_canada\_user\_rating\_pivot1.csv",

encoding = "ISO-8859-1", index\_col='Book-Title')

@app.route('/')# route to display the home page

def home():

return render\_template('home.html')#rendering the home page

@app.route('/extractor')

def extractor():

return render\_template('extractor.html')

#extractor page

@app.route('/keywords', methods=['POST'])# route to show the predictions in a web UI

def keywords():

typ=request.form['type']

output=request.form['output']

#if typ=="text":

#output=re.sub("[^a-zA-Z.,]"," ",output)

print(output)

distances,indices=model.kneighbors(

us\_canada\_user\_rating\_pivot.loc[output,:].values.reshape(1, -1), n\_neighbors = 6)

keyword=[]

for i in range(0, len(distances.flatten())):

if i == 0:

print('Recommendations for {0}:\n'.format(output))

else:

keyword.append('{0}: {1}'.format(i,

us\_canada\_user\_rating\_pivot.index[indices.flatten()[i]]))

# showing the prediction results in a UI

return render\_template('keywords.html',keywords=keywords)

if \_\_name\_\_ == "\_\_main\_\_":

# running the app

app.run(debug=False)