Lungs Cancer Prediction

Abstract

In the human body, the lungs play a very important role. All the cells in the human body needed oxygen and the lungs provide the purified oxygen to all the cells. It takes the needed oxygen into the body and gives all the oxygen to the cells and takes out all the extra carbon dioxide. Lungs work in the system that it takes the desired oxygen in and purifies it with its cleaning system which has small hairs lines used to line the bronchial tube in the lungs having mucus, which is a defense system in the lungs used to move the dirt particles or bacteria and others out of the body. Because the fact that lungs are sensitive and important organs.

In the human body, the lungs are continuously attracted to different types of bacteria and diseases in the air and it made an important effect on the lungs and causes dangerous lung diseases or furthermore can cause lung cancer. The cancer of the lung is caused because of continuous damage occurring to the lung cells, which will cause the disease known as Bronchogenic Carcinoma. This disease will be caused by the rapid spread out of the damaged lung cells. Lung cancer is a top-ranked disease that causes a lot of deaths in cancer-related cases. So, it is vital to make such a system that will predict lung cancer in its early stages and save lives.

Chapter 1 Introduction

1.1. Introduction

In the human body, the lungs play a very important role. All the cells in the human body needed oxygen and the lungs provide the purified oxygen to all the cells. It takes the needed oxygen into the body and gives all the oxygen to the cells and takes out all the extra carbon dioxide. Lungs work in a system that takes the desired oxygen in and purifies it with its cleaning system which has small hairs lines used to line the bronchial tube in the lungs having mucus, which is a defense system in the lungs used to move the dirt particles or bacteria .etc. out of the body. Because the fact that lungs are sensitive and important organs.

In the human body, the lungs are continuously attracted to different types of bacteria and diseases in the air and it made an important effect on the lungs and causes dangerous lung diseases or furthermore can cause lung cancer. Cancer of the lung is caused because of continuous damage occurring to the lung cells, which will cause the disease known as Bronchogenic Carcinoma. This disease will be caused by the rapid spread out of the damaged lung cells. Lung cancer is a top-ranked disease that causes a lot of deaths in cancer-related cases. So, it is vital to make such a system that will predict lung cancer in its early stages and save lives.

As we all know that air pollution, smoking, etc. are the biggest causes of lung cancer. But researcher believes that there is an element known as arsenic, and if its amount gets accusive in the water it can also cause serious lung cancer. The other one is a supplement that increases the speed of diseases in the human body known as beta-carotene. there are empirical pieces of evidence that arsenic in water and beta-carotene supplements also increase the predisposition to the disease. So, we will develop a Lung cancer prediction mechanism using multi-gene genetic programming by selecting automatic features from molecules.

1.2. Overall Explanation

The main objectives of the projects are explained below:

1.2.1. Objectives

The objectives of this project are:

- To build such a methodology which will automatically detect lung cancer in its early stages.
- The build model will use the mutated genes that will reveal the useful information about lung cancer in its early stages in molecules.

- To reduce the death ratio caused by lung cancer by making this kind of system which is more accurate, reliable and time saver.

1.2.2. Problem Description

In past we don't have the technology advance enough to detect the lung cancer in its early stages, because, these past techniques use physical structures of the cancer starts to appears after it took fully control on victims. So, the mortality-rate increases and became the foremost cancer related mortality-rate. So, it is the need of the day to build a framework which will detect lungs cancer in its early stages more accurately and efficiently and we intend to build this system in this project.

1.2.3. Methodology

In every living thing body has millions of cells in it and these cells make sure that the body is alive. Each of these cells in our body contains some information in it and DNA is used to carry this info, we can say that it is equal to a code which is correspond to 1D character string. And these 3 are equal to 1 amino acid. This system is work like parallel because the DNA instructions are translated to- RNA and then protein sequence amino acid and when there is a lung cancer starts causing the cell info got changed and this change will be showed in the amino acids. So that's how we detect cancer in its early stages. And this is how one sample is made. We will input these protein sequences then we will apply preprocessing steps on the data then we will apply feature space formulations and then the model will be applied and at the end it will detect lung cancer.

1.2.4. Product Scope

The system which we are building is evolutionary and make an impact in the medical field. This system will predict the lung cancer in its early stages as compared with the past frameworks, they are not able to do that. The system will help in saving a lot of life's with accuracy and more reliability.

1.2.5. System Requirements

The system will operate on the following specifications

- **Least System Needs**
  - **Operating System:** Windows
  - **CPU Type:** Third Generation of Intel or batter
  - **Storage Needs:** 8 giga bytes

1.3. Expectations and Needs

The system will use molecules dataset which contains different types of amino acid information. Then a revolutionary framework will be applied to detect and predict the lungs cancer and perform some analysis on it.

1.4. External System Necessities

The external system necessities are follow below:

1.4.1. Hardware

- PCs, Laptop

1.4.2. Software

Lung cancer prediction will be build using molecules. Below is the software compliers we will be using in this project.

- Anaconda 3 2022.4(64 bit)
- Jupiter Notebook
- VS Code

1.5. System Characteristics

The system is based on latest AI- based robust programming known as multi-gene genetic programming. This method is ideal for our system because it can automatically detect lung cancer more accurately in its early stages. This method works more efficiently then the past methods.

1.5.1. Response Sequences

When molecules information will be given to the system it will process it and will shows the result by prediction and detection of lung cancer.

Chapter 2 Literature Review

2.1. Introduction

Many research has been done in the past in the analysis of lung cancer but due to incompetent framework they are unable to detect this cancer on time. Basically, lung cancer is very dangerous disease because it is very difficult to identify it and these methods use physical machine to identify. Whereas as different AI-based ML techniques are also applied and analyze the results. And different multiple AI-based methodology is joined with different algos and use image processing techniques to detect lung cancer. But researches have shown that the physical method of detecting the lung cancer is more accurate than these AI and image processing based algos.

Lungs work in the system that it takes the desired oxygen in and purifies it with its cleaning system which has small hairs lines used to line the bronchial tube in the lungs having mucus, which is a defense system in the lungs used to move the dirt particles or bacteria etc. out of the body. Because the fact that lungs are sensitive and important organs. lungs are continuously attracted to different types of bacteria and diseases in the air and it made an important effect on the lungs and causes dangerous lung diseases or furthermore can cause lung cancer.

The cancer of the lung is caused because of continuous damage occurring to the lung cells, which will cause the disease known as Bronchogenic Carcinoma. This disease will be caused by the rapid spread out of the damaged lung cells. Lung cancer is a top-ranked disease that causes a lot of deaths in cancer-related cases. Many solutions of this evil disease are built but they are incompetent systems not able to give the right results. We will build the batter system and solution for this disease which will be able to detect the cancer more accurately.

2.2. Related Works

Many research have been made to tackle the problem in which some of the researcher’s uses different image processing technique with lungs cancer image data to detect lung cancer. Some uses AI-based ML methods to delt with the problem and some uses AI-based DL methods to delt with the problem. The ML method with a physical tool is built in the past to detect lung cancer. A survey is also performed for the two different years to check which has greater rate of this disease and to analyze the air collected as a sample form they use this physical tool.

In the human body, the lungs play a very important role. All the cells in the human body needed oxygen and the lungs provide the purified oxygen to all the cells. It takes the needed oxygen into the body and gives all the oxygen to the cells and takes out all the extra carbon dioxide. Lungs work in the system that it takes the desired oxygen in and purifies it with its cleaning system which has small hairs lines used to line the bronchial tube in the lungs having mucus, which is a defense system in the lungs used to move the dirt particles or bacteria. out of the body. Because the fact that lungs are sensitive and important organ.

The above study shows that if AI-based ML methods embedded with a physical tool shows more accuracy then other methods. Later on a researcher update AI-based algo ANN for the purposes of detecting lung cancer using specific signs. This method is build using a surveyed dataset and it gives the accuracy of above 90%. Another way of detecting lungs cancer is that the diagnosed patients are helping the doctors as they are diagnosing. In this way the doctor gain help from many interpreters at once which will help gaining more accuracy. Many frameworks are applied to tackle this dangerous problem in the past and some of these frameworks gain significant accuracy but there is need of more research in this part of problem.

2.2.1. Terms

The system uses the following terminologies:

- **Support Vector Machine**
- **K-Nearest Neighbors**
- **Decision Tree**
- **Domain:** Artificial Intelligence
- **Dataset**: Dataset of Molecules

2.2.2. Categorization of Existing Techniques/Works/Research

Many research has been done in the past in the analysis of lung cancer but due to incompetent framework they are unable to detect this cancer on time. Basically, lung cancer is very dangerous disease because it is very difficult to identify it but our technique is evolutionary and we aim to detect the lungs cancer with its early stages.

2.2.3. Proposed Improvements in Existing Works

As compared to the past methods our technique will be able to detect lung cancer in its early stages and with high accuracy and reliability. The past methods are mostly use the physical components use for lung cancer detection but it can’t be able to detect the cancer in its early stages. This system will be able to diagnose the cancer in its early stages and able save more lives.

2.3. Summary

Many research has been made to tackle the problem in which some of the researchers uses different image processing technique with lungs cancer image data to detect lung cancer. Some uses AI-based ML methods to delt with the problem and some uses AI-based DL methods to delt with the problem. But this system will use multi-gene genetic programming algo to detect the cancer. Dataset of Molecules is used to train and test the frameworks. This project has what it takes to detect the lungs cancer in its early stages and this way it will save lives of more people and as a result of that the death toll in the world due to lung cancer will be reduced.

2.4. Reference

- Mohsin Sattar, Nabeela Kausar, " Lung cancer prediction using multi-gene genetic programming by selecting automatic features from amino acid sequences".
- Ekmekji, " A Study on Comparison of Lung Cancer Prediction Using Ensemble Machine Learning".
- d. Silva, "Age and Gender Classification – A Proposed System," Department of Computer Science and Technology, Canada.

Chapter 3 System Design

3.1. Introduction

In the current chapter we usually talk about the design framework of the system but this project known as Lung's cancer detection is a researched-based project in which we will use different ML techniques with molecules dataset to detect the lung cancer in its early stages. Our method will automatically detect and predict with the high accuracy rate. As mentioned above that this is a researched base project so in this chapter, we have very little to talk about.

3.1.1. Purpose

We will define some road map for the project in the chapter by creating some activity diagram and workflow diagram for our system which will help in understanding the methodology of the project more easily.

3.1.2. System Overview

The system which we are building is evolutionary and make an impact in the medical field. This system will predict the lung cancer in its early stages as compared with the past frameworks, they are not able to do that. The system will help in saving a lot of lives with accuracy and more reliability.

3.1.3. Design Map

We will input the protein sequences to the system then we will apply preprocessing steps on the data then we will apply feature space formulations on it and then the model will be applied and at the end it will detect lung cancer. We are using multi gene genetic programing for the detection purposes in this project.

3.2. Design deliberation

Because the project is research-based and no UI so there is no discussion about the design is needed.

3.2.1. Expectations

The system use molecules as dataset and genetic programming will apply to detect or predict the lungs cancer.

3.2.2. Restrictions

The system would only accept the data in the form of molecules values other form would not be acceptable.

3.2.3. Hazards and Dangerous Parts

The system is safe and risk-free and it is for the medical field, the project does not have any risks and volatile areas infect the project's aims for public safety.

3.3. Activity Diagram

Figure 3.1 shows the methodology of the system. It shows our system frame of work and the detection process of the system.

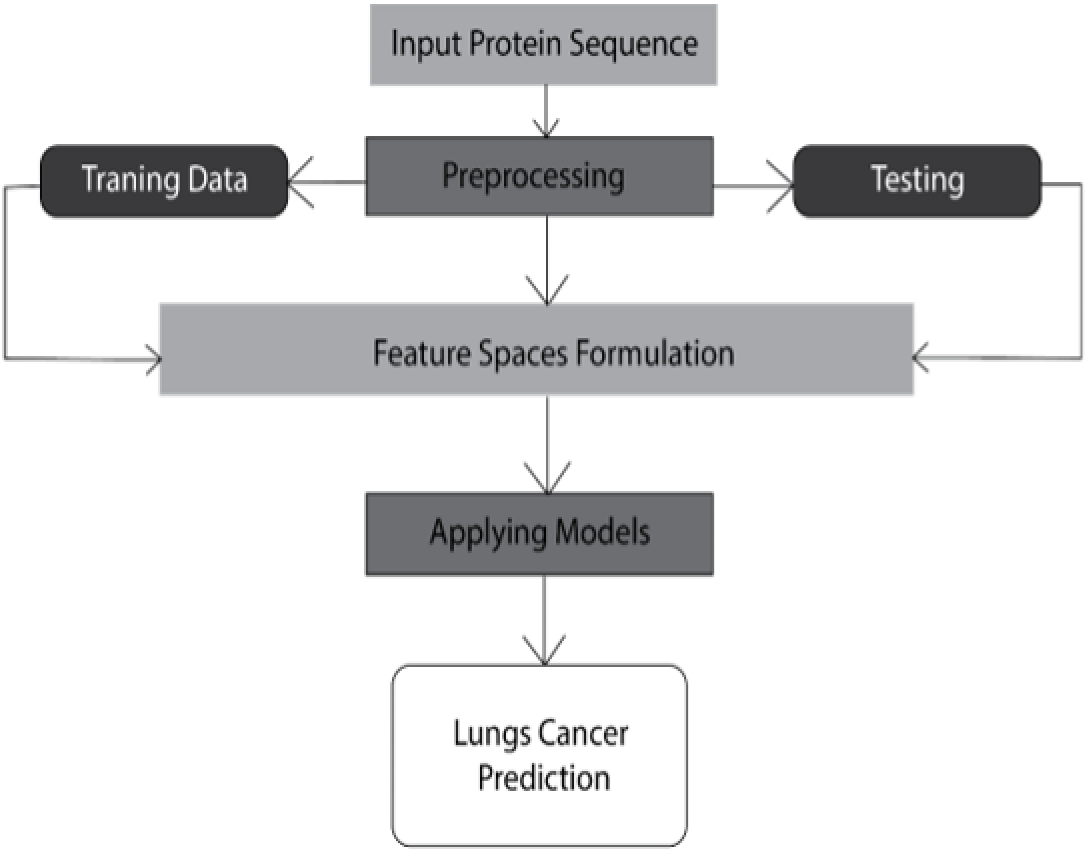


Figure 3. 1 Activity Diagram

This involves collecting a dataset of molecules and associated labels (indicating whether or not each molecule contains lung cancer). Preprocessing may include tasks such as cleaning the dataset and deleting the junk data. Then next step is train a machine learning model. This involves using the training data to train a machine learning model to recognize patterns in the molecules that are indicative of lung cancer. Then we will evaluate the model. Once the model has been trained, it can be evaluated on the validation set to see how well it performs. This may involve calculating metrics such as accuracy, precision, and recall. Then test the model, once the model has been fine-tuned, it can be tested on the test set to get an idea of how well it will generalize to new data.

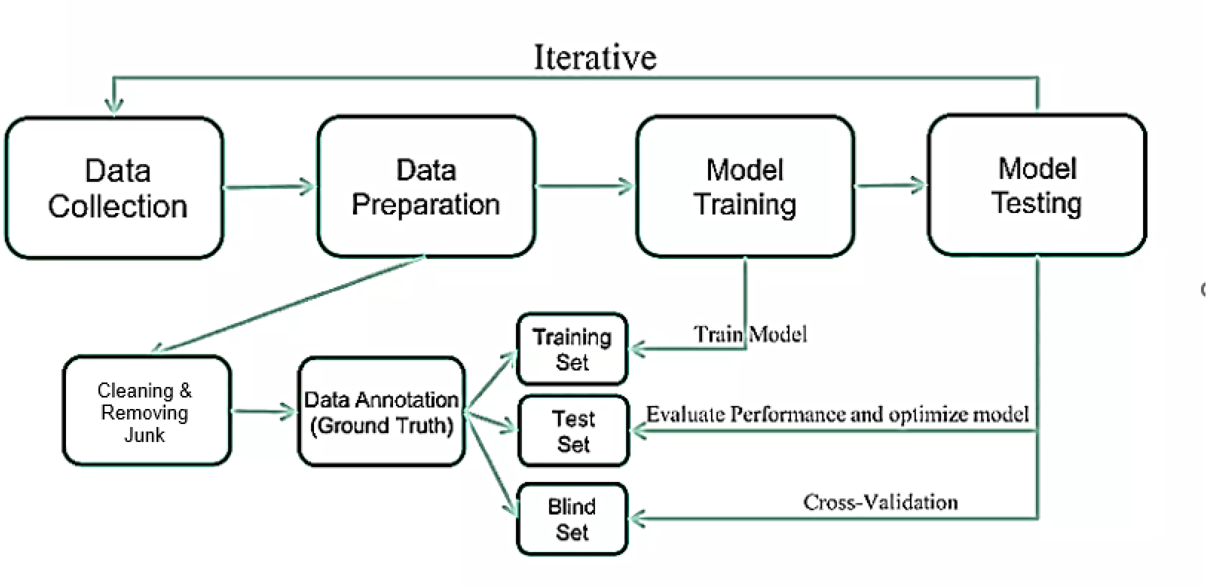


Figure 3. 2 Dataflow Diagram

Figure 3.3 is a data flow diagram (DFD) of lung cancer prediction. It is a graphical representation of the flow of data through a system. It is explained in the following steps:

- Molecules Data is collected from various sources, such as medical records, biopsy results, etc.
- The data is cleaned and preprocessed to prepare it for use in machine-learning models.
- The machine learning models are trained on the preprocessed data.
- The trained model is used to make predictions on new data.
- The predictions are analyzed and reviewed by medical professionals to determine the likelihood of lung cancer.

Chapter 4 Proposed Methodology

4.1. Discussion

In the human body, the lungs are continuously attracted to different types of bacteria and diseases in the air and it made an important effect on the lungs and causes dangerous lung diseases or furthermore can cause lung cancer. The cancer of the lung is caused because of continuous damage occurring to the lung cells, which will cause the disease known as Bronchogenic Carcinoma. This disease will be caused by the rapid spread out of the damaged lung cells. Lung cancer is a top-ranked disease that causes a lot of deaths in cancer-related cases. So, it is vital to make such a system that will predict lung cancer in its early stages and save lives.

4.2. Development Methodologies

In every living thing body has millions of cells in it and these cells make sure that the body is alive. Each of these cells in our body contains some information in it and DNA is used to carry this info, we can say that it is equal to a code which is correspond to 1D character string. And these 3 are equal to 1 amino acid.

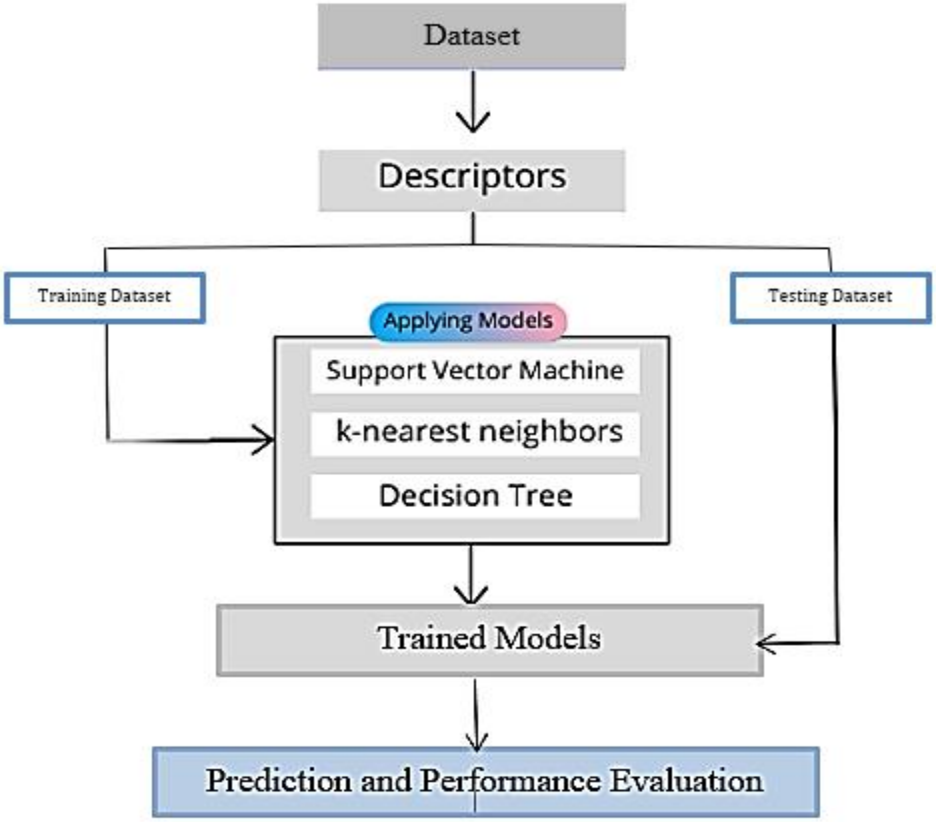


Figure 4. 1 Development Methodology

This system is work like parallel because the DNA instructions are translated to- RNA and then protein sequence amino acid and when there is a lung cancer starts causing the cell info got changed and this change will be showed in the molecules. So that’s how we detect cancer in its early stages. And this is how one sample is made. We will input these molecules samples then we will apply preprocessing steps on the data then we will apply descriptors to extract features and then all the models will be applied and at the end it will detect lung cancer.

The above figure 4.1 shows how the system will work and all the steps to be performed to design such a system. The system first takes a protein sequence as an input and then we perform some preprocessing steps on these protein sequences we use descriptors to extract features and after that, we apply the models to the system. These models are SVM, K-nearest neighbors, and decision tree, then we trained those models and then perform testing on them to predict results.

4.3. Deployment Environment

To build this system we have studied and gain experience and expertise needed for the completion of this project. We have also studies in the relevant field of AI, DL, ML which help us doing this project we will use anaconda as it compiler software and python as programming language to build this project.

4.3.1. Technologies

The following are the technologies used in this project:

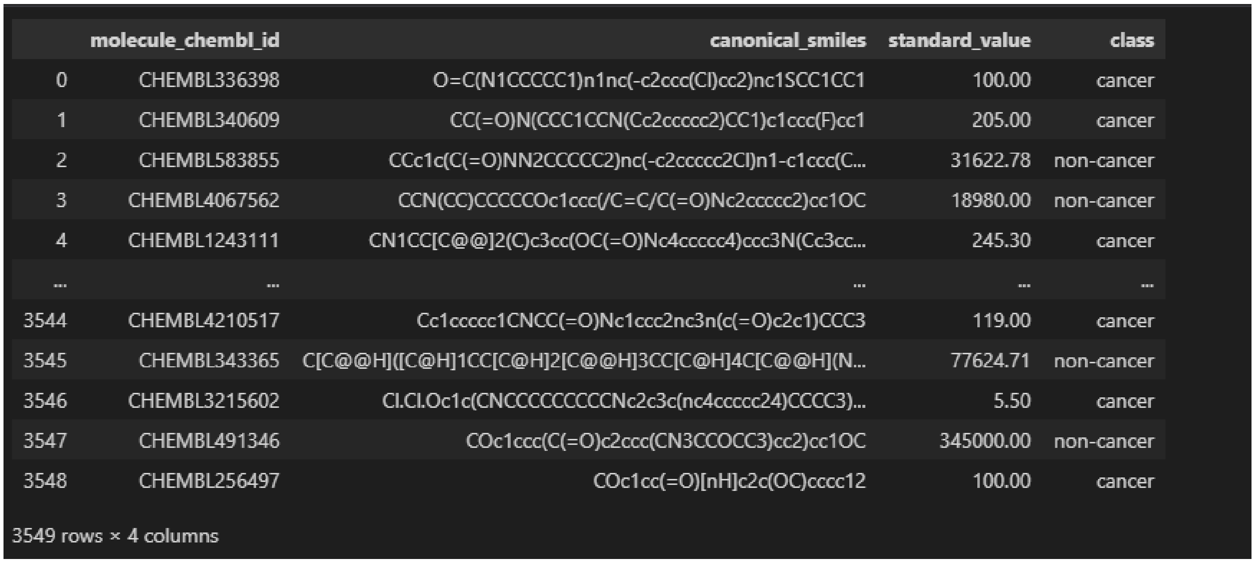
- **Hardware**
  - Personal Laptops
- **Algorithms**
  - Support Vector Machine
  - Decision Tree
  - K-Nearest Neighbors
- **Software**
  - Python
  - Anaconda as IDE.

4.4. Implementation

We don’t have the technology advance enough to detect lung cancer in its early stages in the past, because, these past techniques use physical structures of cancer that start to appear after it took full control of victims. So, the mortality rate increased and became the foremost cancer-related mortality rate. So, it is the need of the day to build a framework that will detect lung cancer in its early stages more accurately and efficiently. In this project, we use the molecules dataset, and then we will clean that dataset, perform some pre-processing on the dataset, and then we apply the descriptors which will extract features, and then models will be applied which will then predict lung cancer.

4.4.1. Dataset

The dataset used for this system is the molecules dataset. This dataset is collected from COSMIC and TCGA datasets. The dataset is divided into two classes, positive and negative and we have above 3000 data samples. The redundant sequences present in the dataset are removed by applying a clustering database of tolerance of above 75% likeness.



	molecule_chembl_id	canonical_smiles	standard_value	class
0	CHEMBL336398	O=C(N1CCCCC1)n1nc(-c2ccc(Cl)cc2)nc1SCC1CC1	100.00	cancer
1	CHEMBL340609	CC(=O)N(CCC1CCN(Cc2ccccc2)CC1)c1ccc(F)cc1	205.00	cancer
2	CHEMBL583855	CCc1c(C(=O)NN2CCCCC2)nc(-c2ccccc2Cl)n1-c1ccc(C...	31622.78	non-cancer
3	CHEMBL4067562	CCN(CC)CCCCCOc1ccc(/C=C/C(=O)Nc2ccccc2)cc1OC	18980.00	non-cancer
4	CHEMBL1243111	CN1CC[C@@]2(C)c3cc(OC(=O)Nc4ccccc4)ccc3N(Cc3cc...	245.30	cancer
...	...	...	...	...
3544	CHEMBL4210517	Cc1ccccc1CNCC(=O)Nc1ccc2nc3n(c(=O)c2c1)CCC3	119.00	cancer
3545	CHEMBL343365	C[C@@H]([C@H]1CC[C@H]2[C@@H]3CC[C@H]4C[C@@H](N...	77624.71	non-cancer
3546	CHEMBL3215602	Cl.Cl.Oc1c(CNCCCCCCCCNc2c3c(nc4ccccc24)CCCC3)...	5.50	cancer
3547	CHEMBL491346	COc1ccc(C(=O)c2ccc(CN3CCOCC3)cc2)cc1OC	345000.00	non-cancer
3548	CHEMBL256497	COc1cc(=O)[nH]c2c(OC)cccc12	100.00	cancer

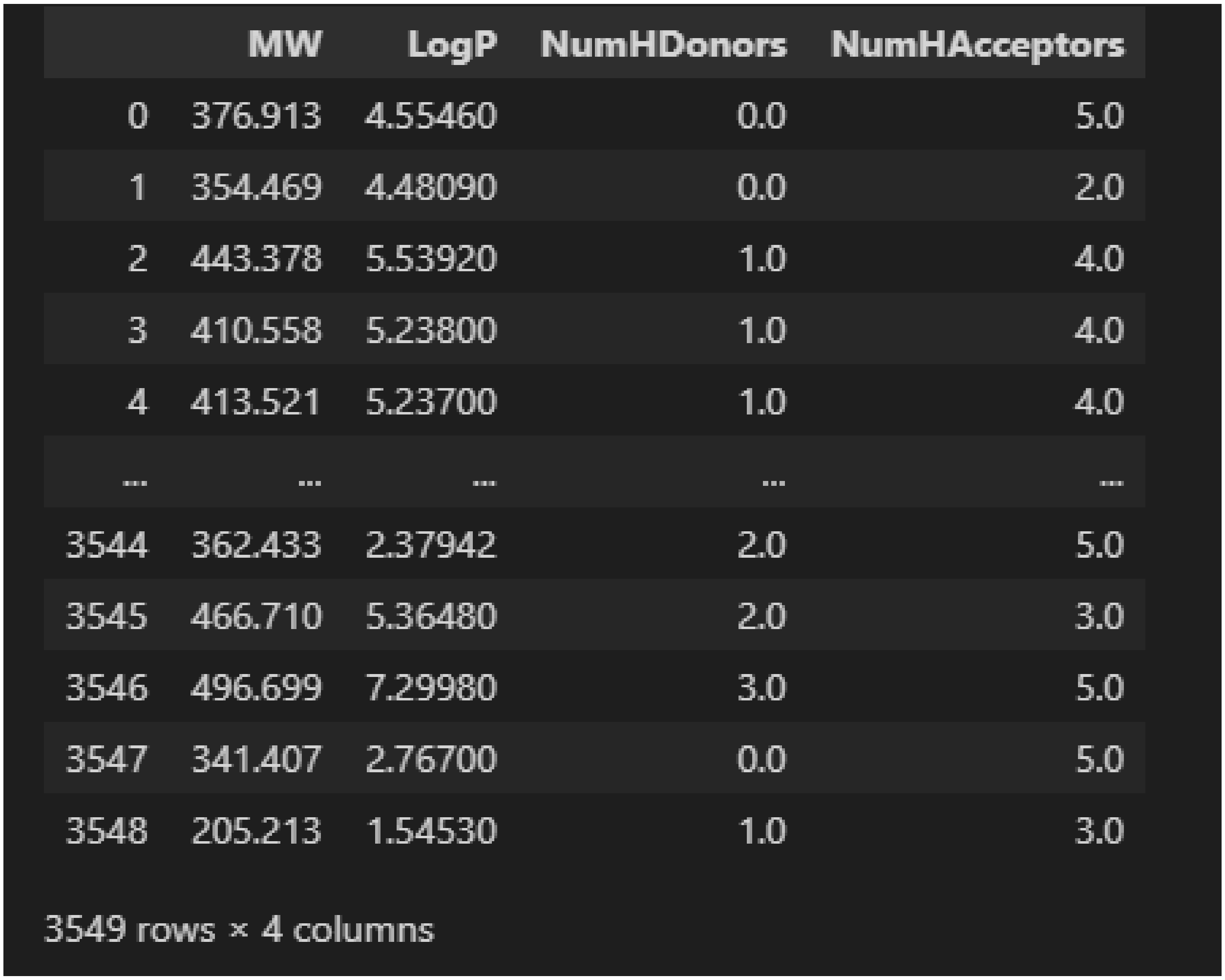
3549 rows × 4 columns

Table 4. 1 Dataset

So now we have decreased the number of examples. Before we perform operations on our dataset will select equal numbers of examples of both classes to form balanced data. We will set 75 % of the dataset for training purposes and the remaining 25% for testing purposes. Table 4.1 shows some of the dataset samples which show that it has 3549 training examples and four columns which include “molecule_chembl id, canonical_smiles, standard_values” with two classes of cancer and not cancer.

4.4.2. Data Preprocessing and Tentative Data Analysis

When the dataset is uploaded to the compiler the next step is to perform some preprocessing steps on the dataset. So firstly we clean the dataset and then we use two descriptors to extract features from the dataset. These descriptors are PaDEL and Lipinski. So Padel and Lipinski will be applied one by one and first, we will apply Lipinski, below in table 4.2 shows the data frame for Lipinski.



	MW	LogP	NumHDonors	NumHAcceptors
0	376.913	4.55460	0.0	5.0
1	354.469	4.48090	0.0	2.0
2	443.378	5.53920	1.0	4.0
3	410.558	5.23800	1.0	4.0
4	413.521	5.23700	1.0	4.0
...	...	...	...	...
3544	362.433	2.37942	2.0	5.0
3545	466.710	5.36480	2.0	3.0
3546	496.699	7.29980	3.0	5.0
3547	341.407	2.76700	0.0	5.0
3548	205.213	1.54530	1.0	3.0

3549 rows × 4 columns

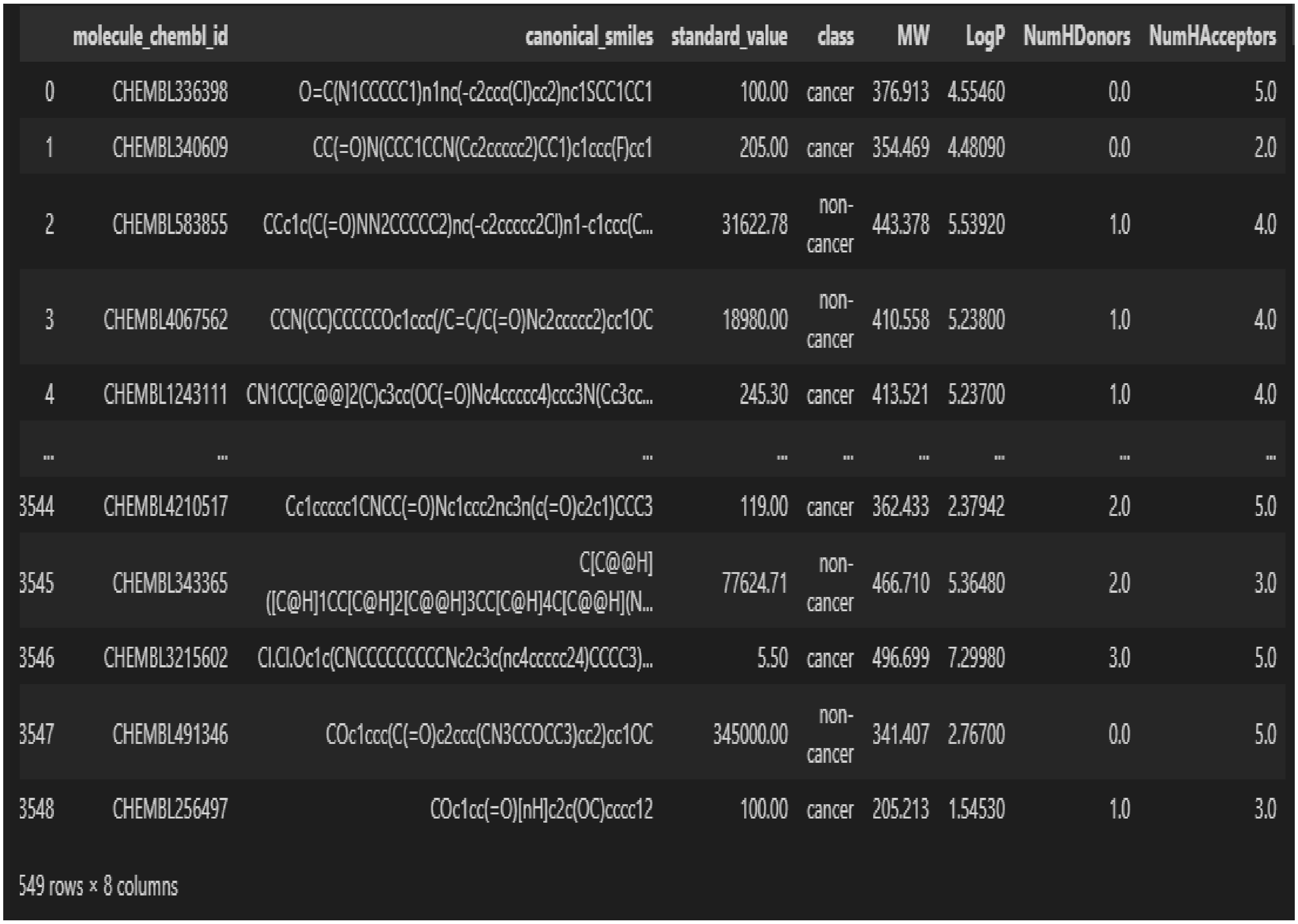
Table 4. 2 Data Frame Lipinski

In above table 4.2, the 1 column is showing the Molecular weight which tells us the size, the second column is the logP and the next one is the relative number of hydrogen bond donors and the last column is of acceptors. The Lipinski descriptors are a set of four quantitative structural parameters that are used to predict the oral bioavailability of a chemical compound. The four Lipinski descriptors are:

1. **Molecular weight**: The weight of a chemical compound, typically measured in grams per mole.

2. **LogP:** The partition coefficient of a chemical compound, which measures its solubility in water and lipids.
3. **Number of hydrogen bond donors:** The number of atoms in the chemical compound that can donate a hydrogen bond.
4. **Number of hydrogen bond acceptors:** The number of atoms in the chemical compound that can accept a hydrogen bond.

Compounds that have high values for these descriptors are more likely to have poor oral bioavailability, meaning they are less likely to be absorbed into the body when taken orally. And now we will combine the data frame of Lipinski with the original data frame as shown in table 4.3.



	molecule_chembl_id	canonical_smiles	standard_value	class	MW	LogP	NumHDonors	NumHAcceptors
0	CHEMBL336398	O=C(N1CCCCC1)n1nc(-c2ccc(Cl)cc2)nc1SCC1CC1	100.00	cancer	376.913	4.55460	0.0	5.0
1	CHEMBL340609	CC(=O)N(CCC1CCN(Cc2ccccc2)CC1)c1ccc(F)cc1	205.00	cancer	354.469	4.48090	0.0	2.0
2	CHEMBL583855	CCc1c(C(=O)NN2CCCCC2)nc(-c2ccccc2Cl)n1-c1ccc(C...	31622.78	non-cancer	443.378	5.53920	1.0	4.0
3	CHEMBL4067562	CCN(CC)CCCCCOc1ccc(/C=C/C(=O)Nc2ccccc2)cc1OC	18980.00	non-cancer	410.558	5.23800	1.0	4.0
4	CHEMBL1243111	CN1CC[C@@]2(C)c3cc(OC(=O)Nc4ccccc4)ccc3N(Cc3cc...	245.30	cancer	413.521	5.23700	1.0	4.0
...	...	...	...	...	...	...	...	...
3544	CHEMBL4210517	Cc1ccccc1CNCC(=O)Nc1ccc2nc3n(c(=O)c2c1)CCC3	119.00	cancer	362.433	2.37942	2.0	5.0
3545	CHEMBL343365	C[C@@H] ([C@H]1CC[C@H]2[C@@H]3CC[C@H]4C[C@@H](N...	77624.71	non-cancer	466.710	5.36480	2.0	3.0
3546	CHEMBL3215602	Cl.Cl.Oc1c(CNCCCCCCCCCNc2c3c(nc4ccccc24)CCCC3)...	5.50	cancer	496.699	7.29980	3.0	5.0
3547	CHEMBL491346	COc1ccc(C(=O)c2ccc(CN3CCOCC3)cc2)cc1OC	345000.00	non-cancer	341.407	2.76700	0.0	5.0
3548	CHEMBL256497	COc1cc(=O)[nH]c2c(OC)cccc12	100.00	cancer	205.213	1.54530	1.0	3.0

549 rows × 8 columns

Table 4. 3 Combined Data Frames

Below in figure 4.1 shows the graph between two classes of lung cancer. This shows that the active cancer classes are a little higher in the frequency ratio as compared to the other class.

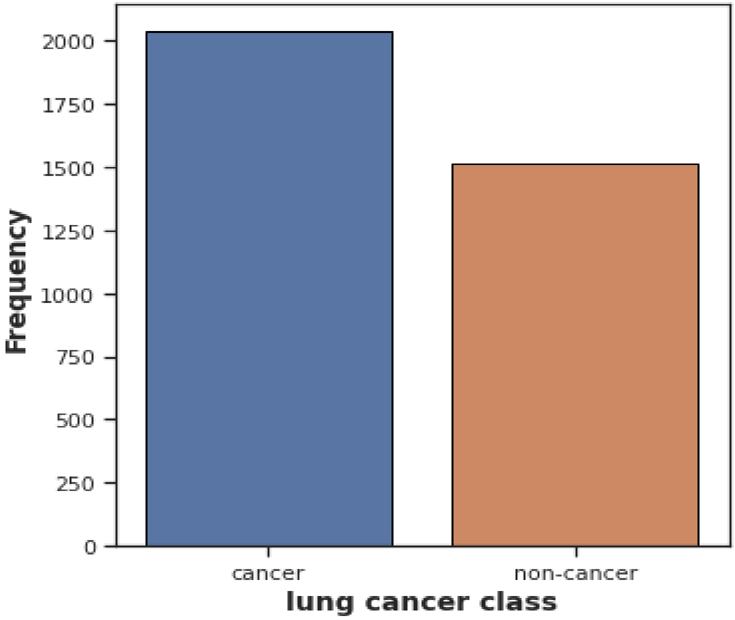


Figure 4. 2 Class x Frequency

Below figure 4.2 is a scatter plot that shows the plot against molecular weight and logP and it shows the variation in the two cancer classes.

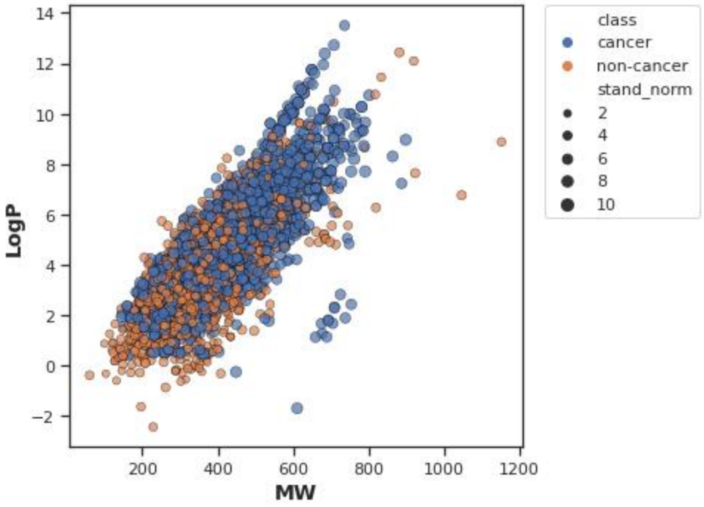
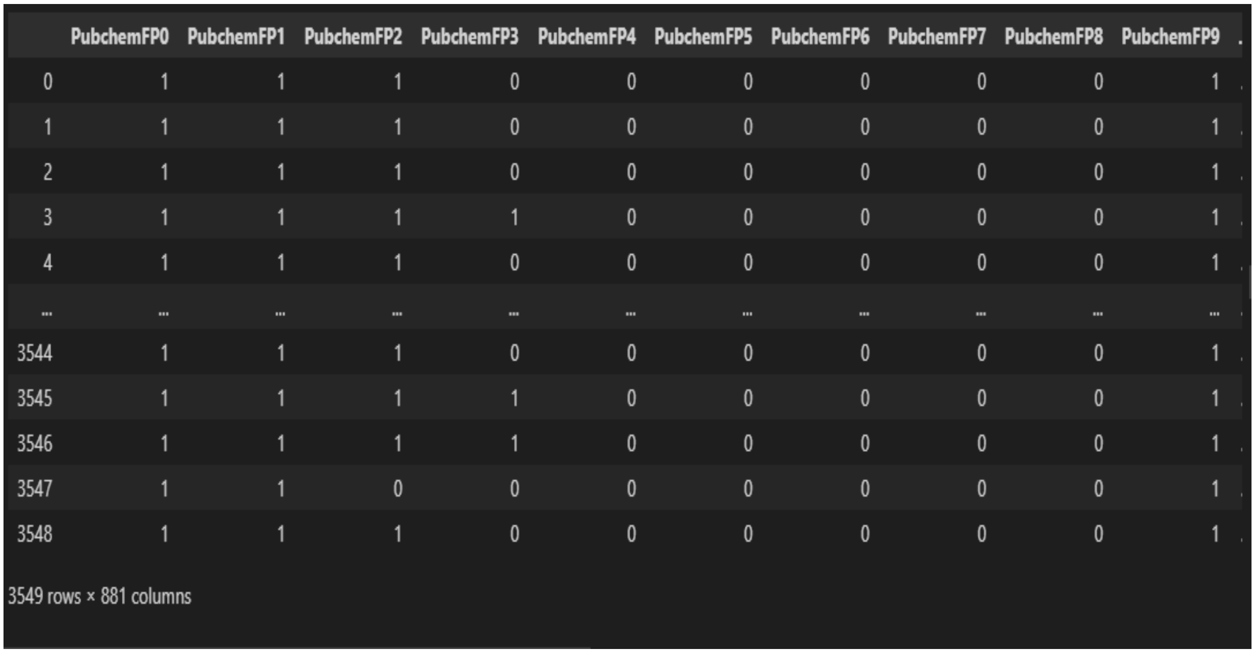


Figure 4. 3 MW x LogP

Now we will apply the second descriptor known as PaDEL in which we will perform the same steps as above performed in the Lipinski. Padel descriptors are a set of numerical chemical descriptors that are used to represent chemical compounds in a way that can be used as input to a machine learning model. Padel descriptors are calculated based on the atomic properties and topological features of a chemical compound. These descriptors can be used as features in a machine learning model to predict various properties of the chemical. There are over 2,000 Padel descriptors that can be calculated for a given chemical compound. These descriptors can be used as features in a machine learning model to predict various properties of the chemical. So, the below in table 4.4 shows the data frame for Padel.



	PubchemFP0	PubchemFP1	PubchemFP2	PubchemFP3	PubchemFP4	PubchemFP5	PubchemFP6	PubchemFP7	PubchemFP8	PubchemFP9
0	1	1	1	0	0	0	0	0	0	1
1	1	1	1	0	0	0	0	0	0	1
2	1	1	1	0	0	0	0	0	0	1
3	1	1	1	1	0	0	0	0	0	1
4	1	1	1	0	0	0	0	0	0	1
...	...	...	...	...	...	...	...	...	...	...
3544	1	1	1	0	0	0	0	0	0	1
3545	1	1	1	1	0	0	0	0	0	1
3546	1	1	1	1	0	0	0	0	0	1
3547	1	1	0	0	0	0	0	0	0	1
3548	1	1	1	0	0	0	0	0	0	1

3549 rows × 881 columns

Table 4. 4 Data Frame PaDEL

So now we will apply the model on the final dataset and train it and then test it.

4.4.3. Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning algorithm used for both classification and regression. Though we say regression problems as well its best suited for classification. The objective of the SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points. The dimension of the hyperplane depends upon the number of features. If the number of input features is two, then the hyperplane is just a line. If the number of input features is three, then the hyperplane becomes a 2-D plane. It becomes difficult to imagine when the number of features exceeds three. Now we will apply the ML model known as the Support Vector Machine. It is a supervised ML algorithm that works well against classification problems. After applying the SVM model the system gives an accuracy of 82%.

4.4.4. Decision Tree

A decision tree is one of the simplest yet highly effective classifications and prediction visual tools used for decision-making. It takes a root problem or situation and explores all the possible scenarios related to it based on numerous decisions. Since decision trees are highly resourceful, they play a crucial role in different sectors. From programming to business analysis, decision tree examples are everywhere. At first, a decision tree appears as a tree-like structure with different nodes and branches.

It is based on the classification principles that predict the outcome of a decision, leading to different branches of a tree. It starts from a root, which gradually has different decision nodes. The structure has terminating nodes in the end. This algorithm works excellently against classification problems. After applying the algorithm, we get an accuracy of 87%.

4.4.5. K-Nearest Neighbors

K-Nearest Neighbors is one of the most basic yet essential classification algorithms in Machine Learning. It belongs to the supervised learning domain and finds intense application in pattern recognition, data mining, and intrusion detection. It is widely disposable in real-life scenarios since it is non-parametric, meaning, it does not make any underlying assumptions about the distribution of data. We are given some prior data (also called training data), which classifies coordinates identified by an attribute. Our last algorithm is K-nearest neighbors. Which is a supervised ML algorithm used for classification problems. After applying the algorithm we get the highest accuracy of 91%.

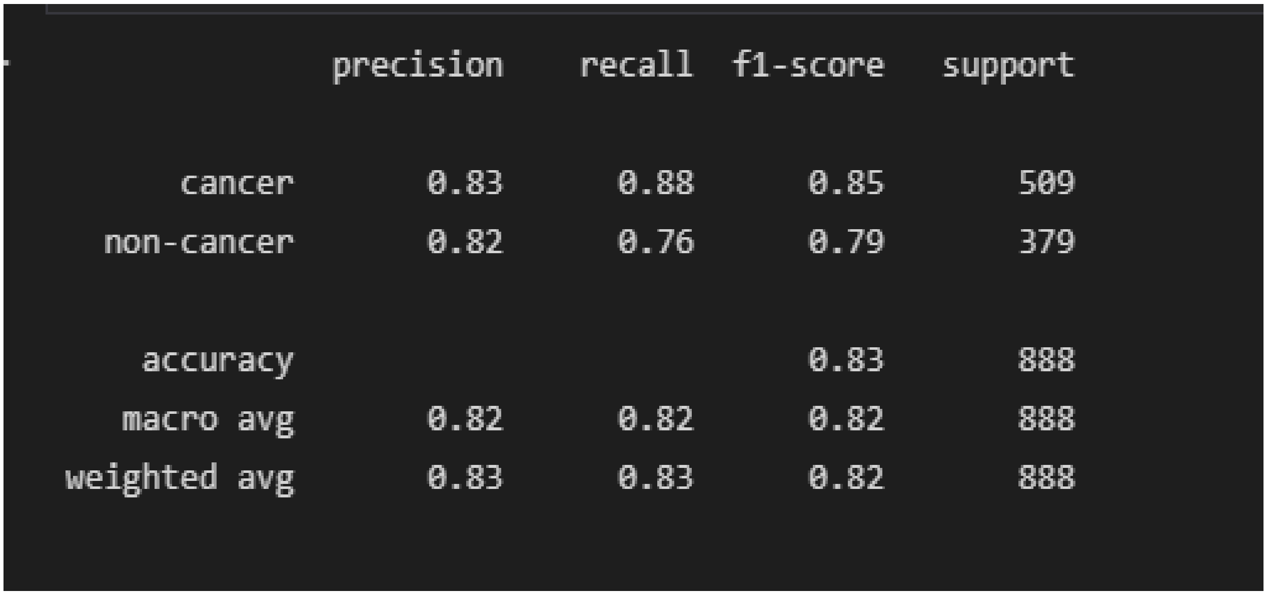
Chapter 5 Results and Discussions

5.1 Discussion

The essential part of any newly designed system is to test it properly and make a proper evaluation of the system to avoid any bugs and errors appearing later in the system when deploying. So we also make some evaluations and perform some tests on our structure to make the ensuratiy of its smooth running in future. These scenarios are applied to certain systems parameters like input and then output predictions.

5.2 System Outputs and Analysis

We have performed different test use cases on our system to check its accuracy and reliability. Table 5.1 show the confusion matrix of the system when applied support vector machine algorithm. It shows the precision-recall and f1-score of the support vector machine algorithm. It shows that the precision, recall and f1-score of the Support vector machine is little higher for the cancer classes as compared to non-cancer classes. When we applied the SVM model the system gives an accuracy of 82%.



```
              precision    recall  f1-score   support

      cancer       0.83      0.88      0.85       509
  non-cancer       0.82      0.76      0.79       379

    accuracy                           0.83       888
   macro avg       0.82      0.82      0.82       888
weighted avg       0.83      0.83      0.82       888
```

Table 5. 1 Confusion Matrix SVM

The figure, 5.1 shows the heatmap which shows the support vector machine confusion matrix result in the visualization form. The dark blue box shows that actual cancer and correctly predicted are more in number than the false predicted class by the algorithm as shown below:

.

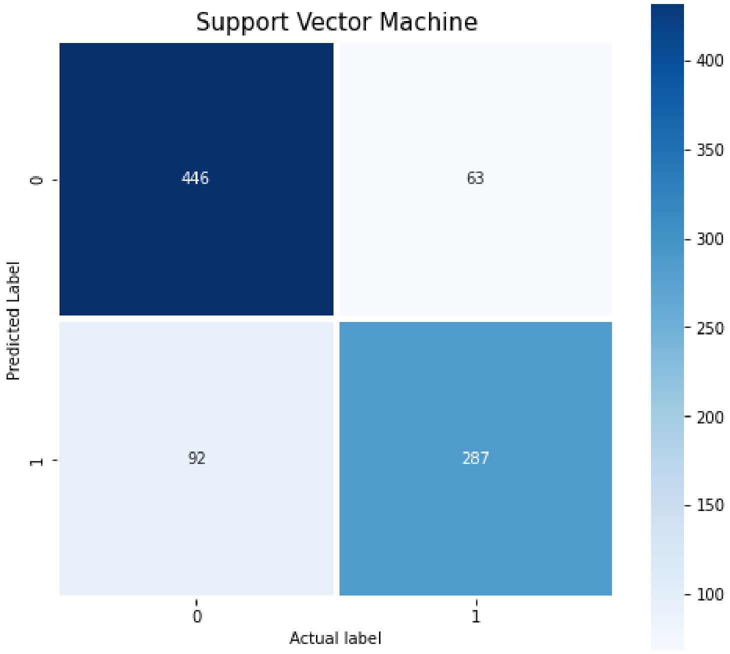
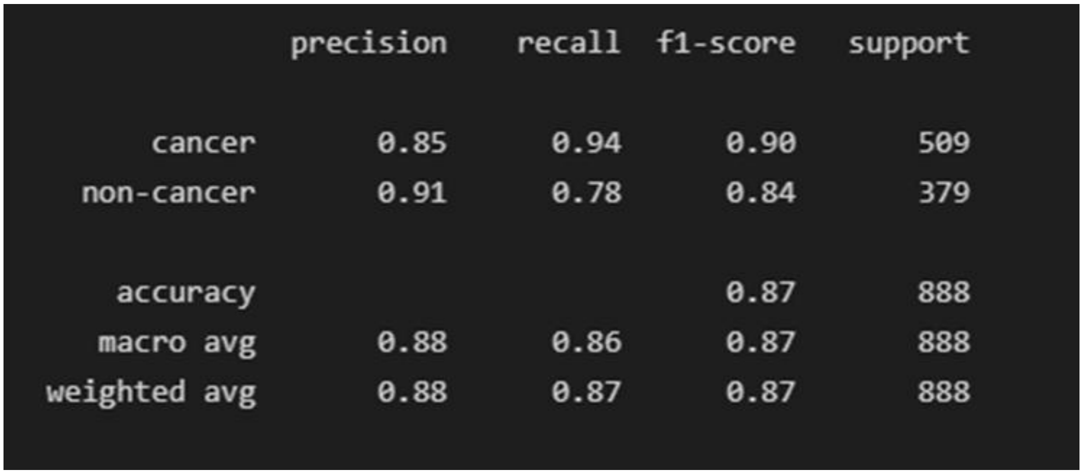


Figure 5. 1 Heatmap SVM

Table 5.2 show the confusion matrix of the system when applied to the Decision Tree algorithm. It shows the precision-recall and f1-score of the Decision tree algorithm. It shows that the precision, recall and f1-score of the decision tree is little higher for the cancer classes in recall and f1-score as compared to non-cancer classes. When we applied the decision tree it gives us an accuracy of 87%.



```
              precision    recall  f1-score   support

      cancer       0.85      0.94      0.90       509
  non-cancer       0.91      0.78      0.84       379

    accuracy                           0.87       888
   macro avg       0.88      0.86      0.87       888
weighted avg       0.88      0.87      0.87       888
```

Table 5. 2 Confusion Matrix Decision Tree

The figure, 5.3 shows the heatmap which shows the Decision tree confusion matrix result in the visualization form. The dark blue box shows that actual cancer and correctly predicted are more in number than the false predicted class by the algorithm as shown below:

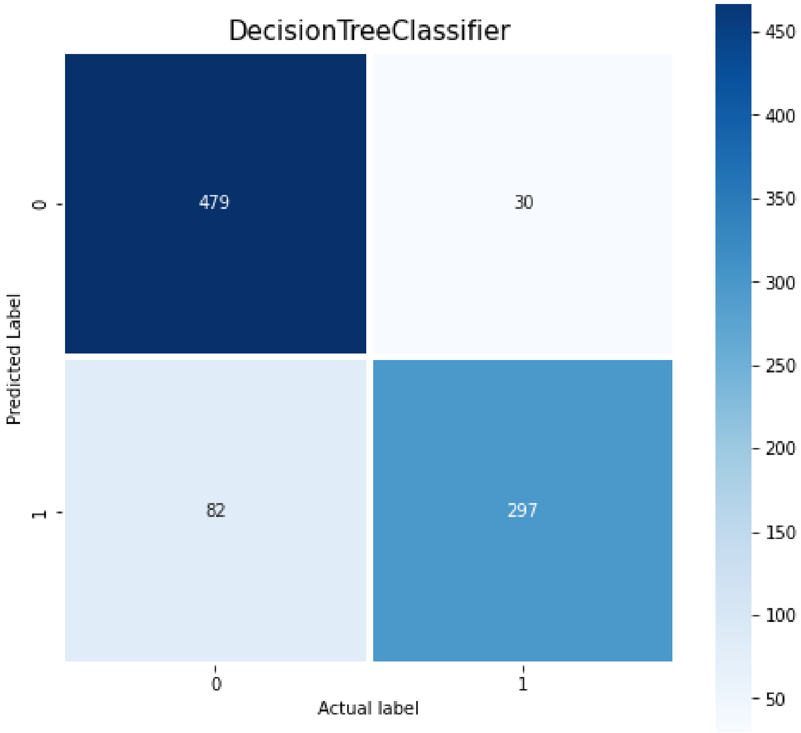
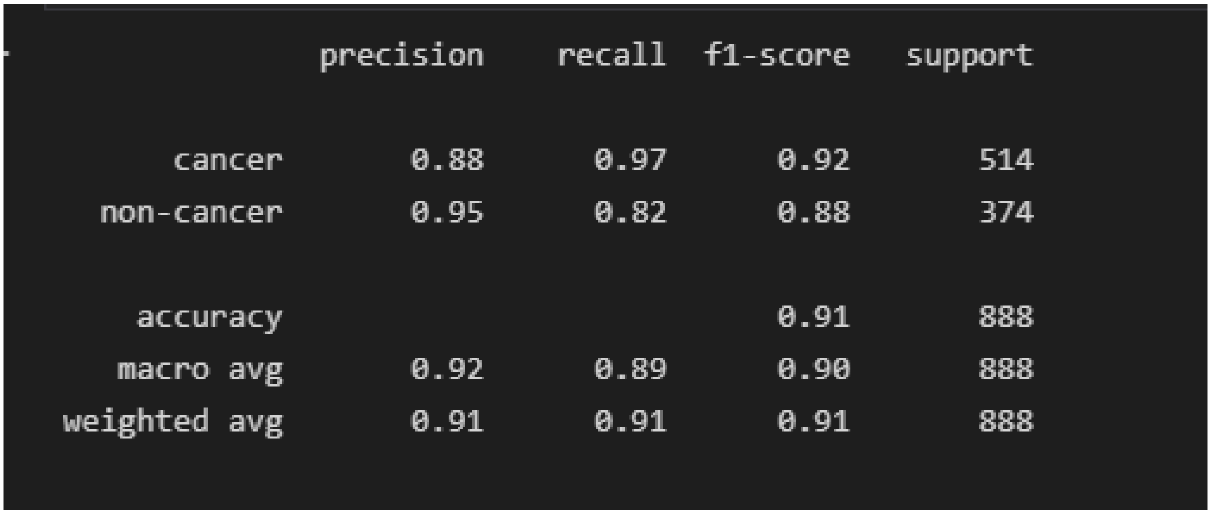


Figure 5. 2 Heatmap Decision Tree

Table 5.3 show the confusion matrix of the system when applied to the K-Nearest Neighbors algorithm. It shows the precision-recall and f1-score of the k-nearest neighbor's algorithm. It shows that the precision, recall and f1-score of the k-nearest neighbors is little higher for the cancer classes in recall and f1-score as compared to non-cancer classes. When we applied the k-nearest neighbors it gives us the highest accuracy of 91%.



```
              precision    recall  f1-score   support

      cancer       0.88      0.97      0.92       514
  non-cancer       0.95      0.82      0.88       374

    accuracy                           0.91       888
   macro avg       0.92      0.89      0.90       888
weighted avg       0.91      0.91      0.91       888
```

Table 5. 3 Confusion Matrix K-Nearest Neighbors

Below figure 5.5 show the Heat map of Confusion matrix of the system when applied K-Nearest Neighbor algorithm. The dark blue box shows that actual cancer and correctly predicted are more in number than the false predicted class by the algorithm.

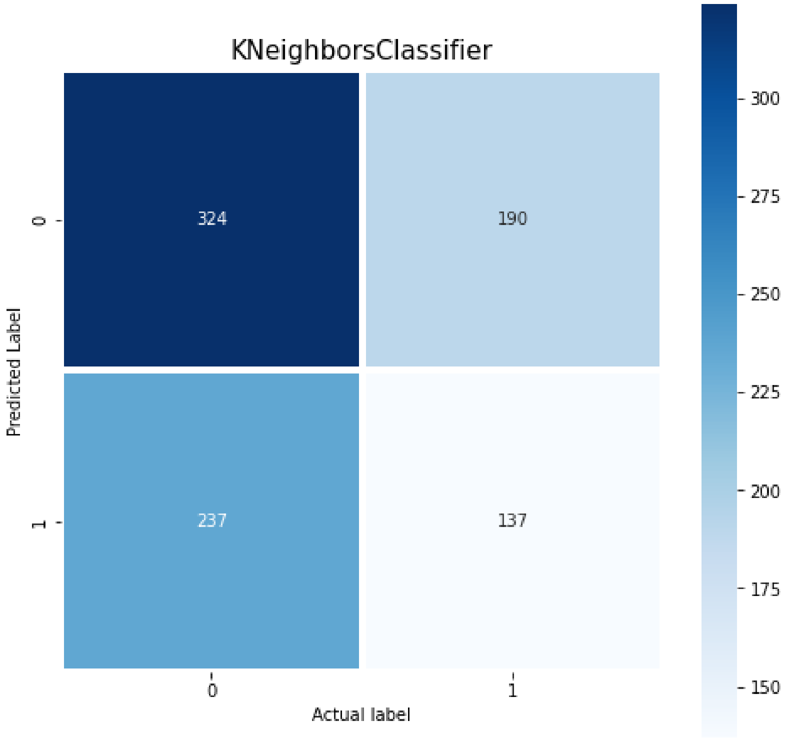


Figure 5. 3 Heatmap K-nearest Neighbors.

.

5.3 Summary

This is the system build to detect lung cancer using AI-based automatic Machine Learning techniques known as SVM, K-Nearest Neighbors, and decision tree. The system uses a protein sequences dataset and with the help of descriptors features extraction of the dataset is done. The system predicts cancer accurately however if we increase the size of the dataset and apply different other models this system will perform more accurately.

.

Chapter 6 Conclusion and Future Work

6.1 Decision

We conclude in this chapter that we can make an accurate, reliable, and user-friendly. It can detect lung cancer using AI-based automatic Machine Learning techniques known as SVM, K-Nearest Neighbors, and decision tree. The system uses a protein sequences dataset and with the help of descriptors features extraction of the dataset is done. The system predicts cancer accurately however if we increase the size of the dataset and apply different other models this system will perform more accurately.

6.2 Upgrades and Updates

With the rapid progress and improvement in technology and most importantly in the field of Artificial Intelligence when every day a new algorithm is developed there is a definite place for an improvement in this system of lung cancer prediction. This advancement unlocks many options and gives us more ground to cover, as we can increase the dataset or change the dataset type like taking image data of lung cancer and using different computer vision techniques and other Deep learning techniques to predict lung cancer. It will be a more accurate, reliable, and flexible system. however, at this time, our system is good and accurately predicts lung cancer. The below figure also show that the future of AI and its growth in all kinds of industries as the technological advancement continues to happen the AI will also rise and the lungs cancer detection can be more improved with this growth.

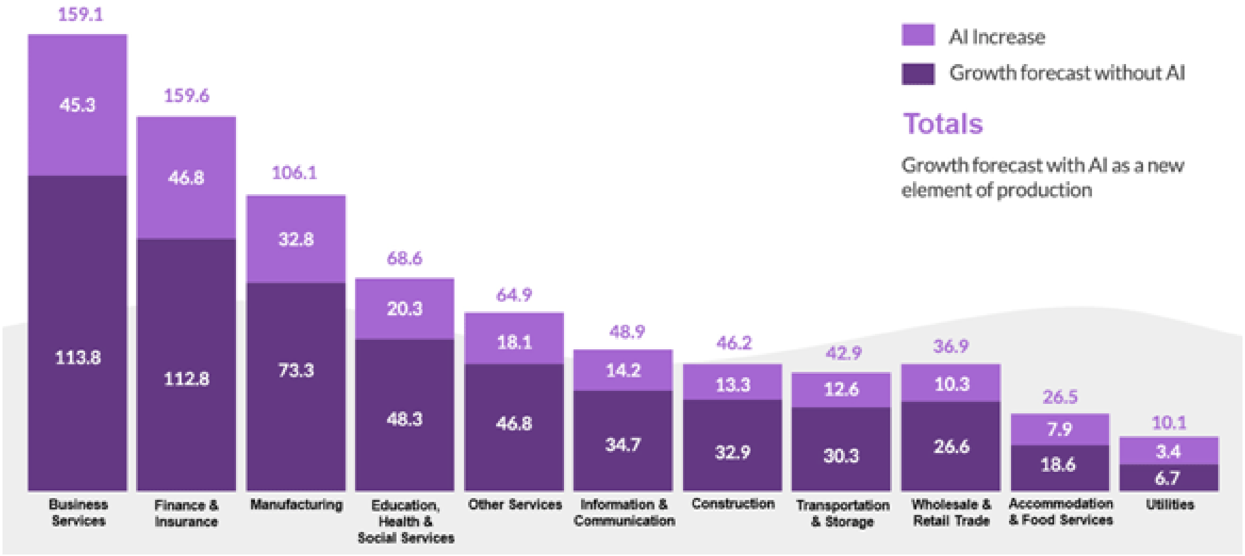


Figure 6. 1 Growth in AI

References

[1] A. Ekmekji, " A Study on Comparison of Lung Cancer Prediction Using Ensemble Machine Learning".

[2] Mohsin Sattar, Nabeela Kausar, " Lung cancer prediction using multi-gene genetic programming by selecting automatic features from amino acid sequences".

[3] " A Review of most Recent Lung Cancer Detection Techniques using Machine Learning".

[4] T. Hassner, " Detection and Prediction of Lung Cancer Using Different Algorithms," The Open University of UK.

[5] Singla, " A Review of Lung cancer Prediction System using Data Mining Techniques and Self Organizing Map (SOM)," Data Science Blogathon.

[6] d. Silva, "Age and Gender Classification – A Proposed System," Department of Computer Science and Technology, Canada.