

# **LUNG-CANCER STAGE PREDICTION USING MACHINE**

## **LEARNING APPROACH**

**BY**

**Md Abdus Salam Patwary**

**ID: 201-15-14004**

The requirements for the Bachelor of Science in Computer Science and Engineering are partially satisfied by this report.

Supervised By

**Md Ali Hossain**

Assistant Professor

Department of CSE

Daffodil International University

Co-Supervised By

**Md Abdus Sattar**

Assistant Professor & Coordinator M.sc

Department of CSE

Daffodil International University



**DAFFODIL INTERNATIONAL UNIVERSITY**

**DHAKA, BANGLADESH**

**January 2024**

## **APPROVAL**

This Project/internship titled “**Lung-Cancer Stage Prediction Using Machine Learning Approach**”, submitted by **Md Abdus Salam Patwary, ID: 201-15-14004** to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 21.01.2024.

### **BOARD OF EXAMINERS**



**Dr. Md. Ismail Jabiullah (MIJ)**

**Professor**

Department of CSE

Faculty of Science & Information Technology

Daffodil International University

**Chairman**



**Raja Tariqul Hasan Tusher (THT)**

**Assistant Professor**

Department of CSE

Faculty of Science & Information Technology

Daffodil International University

**Internal Examiner 1**



**Mayen Uddin Mojumdar (MUM)**

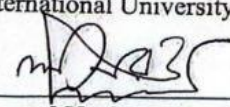
**Senior Lecturer**

Department of CSE

Faculty of Science & Information Technology

Daffodil International University

**Internal Examiner 2**



**Dr. Mohammad Nasir Uddin (DNU)**

**Professor**

Department of Computer Science and Engineering

Jagannath University

**External Examiner 1**


©Daffodil International university

iii

## DECLARATION

We hereby declare that this project has been done by us under the supervision of **Md. Ali Hossain, Assistant Professor**, Department of CSE Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

**Supervised by:**



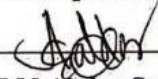
**Md Ali Hossain**

**Assistant Professor**

Department of CSE

Daffodil International University

**Co-Supervised by:**



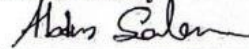
**Md Abdus Sattar**

**Assistant Professor & Coordinator M.sc**

Department of CSE

Daffodil International University

**Submitted by:**



**Md Abdus Salam Patwary**

ID: 201-15-14004

Department of CSE

Daffodil International University

## ACKNOWLEDGEMENT

First of all, allow me to sincerely thank and feel glad to the Almighty Allah for His heavenly favour, which enabled us to successfully finish the final year Thesis based project.

I sincerely thank **Md. Ali Hossain, Assistant Professor, CSE Department**, Daffodil International University, Ashulia, Savar, and send my best wishes. My supervisor is really interested in using machine learning to do this work since he has a thorough grasp of it. His endless patience, guidance in the classroom, unwavering support, intense oversight, constructive criticism, perceptive advice, reading several poorly done drafts, and meticulous editing made it possible to complete this task.

Sincere gratitude is extended to the Professor **Dr. Sheak Rashed Haider Noori** head of the CSE department of Daffodil International University. As well as the other faculty members and staff for their indispensable support in seeing my project through to completion.

I want to express my gratitude to all of my Daffodil International University students who took part in this conversation while still in class.

Lastly, I must respectfully thank my parents for their unwavering support and patience.

## **ABSTRACT**

Physical illnesses like lung cancer have become more common nowadays. The world nowadays is aware of the subject. The disease known as lung cancer affects most individuals. A measure of the sickness is the variations in diagnostic report ratios between patients who are normal and those who are afflicted. Numerous investigations have already been conducted on the illness of lung cancer. I've identified a few excellent chances to improve the procedure even further. I suggest using effective algorithm models to anticipate hazards and raise early alert. My suggested approach is easy to utilize in practice and appropriate for basic projections of lung cancer sickness. The GitHub website served as the dataset's host. Several five distinct kinds of algorithms, including Decision Trees, KNN, ANN, SVM, and LR (logistic regression), have been used. In order to defend the performances, I additionally used a few other ensemble models. ANN had the accuracy at 99.50%, KNN at 99.50%, and SVM at around 97.50%. After that, I received a perfect score of 100% in both Decision Tree and Logistic Regression. I optimized each classifier's parameters using hyper-parameter tweaking. The experimental research analyzed the findings of other recent studies and produced more accurate estimates of lung cancer disease, with 100% accuracy being the greatest performance.

**Keywords: Lung Cancer, K-Neighbors Classifier, Correlation, Prediction, Machine Learning, Algorithms**

## TABLE OF CONTENTS

<b>CONTENTS</b>	<b>PAGE</b>
Board of examiners	ii
Declaration	iii
Acknowledgements	iv
Abstract	v
<b>CHAPTER</b>	
<b>CHAPTER 1: INTRODUCTION</b>	<b>1-4</b>
1.1 Introduction	1
1.2 Motivation	2
1.3 Rationale of the Study	2
1.4 Research Questions	3
1.5 Expected Output	3
1.6 Project Management and Finance	3
1.7 Report Layout	4

<b>CHAPTER 2: BACKGROUND</b>	<b>5-7</b>
2.1 Preliminaries/Terminologies	5
2.2 Related Works	5
2.3 Comparative Analysis and Summary	6
2.4 Scope of the Problem	7
2.5 Challenges	7
<b>CHAPTER 3: RESEARCH METHODOLOGY</b>	<b>8-13</b>
3.1 Research Subject and Instrumentation	8
3.2 Data Collection Procedure	8
3.3 Statistical Analysis	11
3.4 Proposed Methodology	12
3.5 Implementation Requirements	13
<b>CHAPTER 4: EXPERIMENTAL RESULTS AND DISCUSSION</b>	<b>14-31</b>
4.1 Experimental Setup	14
4.2 Experimental Results & Analysis	18

4.3 Discussion	30
<b>CHAPTER 5: IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY</b>	<b>32-33</b>
5.1 Impact on Society	32
5.2 Impact on Environment	32
5.3 Ethical Aspects	33
5.4 Sustainability Plan	33
<b>CHAPTER 6: SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH</b>	<b>34-35</b>
6.1 Summary of the Study	34
6.2 Conclusions	34
6.3 Implication for Further Study	34
6.4 Limitations	35
<b>REFERENCES</b>	<b>36-39</b>
<b>PLAGIARISM REPORT</b>	<b>40</b>



## LIST OF FIGURES

<b>FIGURES</b>	<b>PAGE NO</b>
Figure 3.1: Number of target values	8
Figure 3.2: Methodology of Lung cancer disease	11
Figure 3.3: Correlated Features of Lung Cancer Dataset	12
Figure 4.1: SVM classifier	14
Figure 4.2: Decision Tree Classifier	15
Figure 4.3: Artificial Neural Network	16
Figure 4.4: K-Nearest	16
Figure 4.5: Logistic Regression Classifier	17
Figure 4.6: Experimental Results of Classifiers	18
Figure 4.7: Confusion Matrix Analysis of Logistic Regression	19
Figure 4.8: Cross Validation Analysis of Logistic Regression	20
Figure 4.9: AUC-ROC Curve Analysis of Logistic Regression	20
Figure 4.10: Confusion Matrix Analysis of Decision Tree	21
Figure 4.11: Cross Validation Analysis of Decision Tree	22
Figure 4.12: AUC-ROC Curve Analysis of Decision Tree	22
Figure 4.13: Confusion Matrix Analysis of ANN	23

Figure 4.14: Cross Validation Analysis of ANN	24
Figure 4.15: AUC-ROC Curve Analysis of Artificial Neural Network	24
Figure 4.16: Confusion Matrix Analysis of KNN	25
Figure 4.17: Cross Validation Analysis of KNN	25
Figure 4.18: AUC-ROC Curve Analysis of KNN	26
Figure 4.19: Confusion Matrix Analysis of SVM	27
Figure 4.20: Cross Validation Analysis of SVM	27
Figure 4.21: AUC-ROC Curve Analysis of SVM	28
Figure 4.22: Compilation Time (Second)	28

## LIST OF TABLES

<b>TABLES</b>	<b>PAGE NO</b>
Table 3.1: Details of the dataset	8
Table 3.2: Comparative Analysis	6

# CHAPTER 1

## INTRODUCTION

### 1.1 Introduction

The worst aspect of our daily lives is lung cancer sickness, which encompasses a range of lung conditions including infection and declining function. This patient is showing up more and more often. However, the main difficulty is in identifying or finding the damaged causes at the time of diagnosis. By examining a range of factors and patient diagnostic data, artificial intelligence (AI) may be the most helpful component of a substantial role in forecasting the chance of lung cancer illness from reactive health information. I have reviewed the patient's diagnostic records throughout our investigation and have determined several important signs of the condition. The dataset focuses on determining a person's status for lung cancer and doing bodily tests. Together with other researchers, they developed machine-learning approaches to recognize the illness in the body. But their methodology and precision were neither straightforward nor appropriate for predicting lung cancer cases. I present my method to improve the accuracy of disease prognosis in the human body. There are two recognized categories for learning strategies. While the other is not being observed, one of them is. Using examples of input-output pairings and labelled data, supervised learning produces outputs from inputs. The operational data comes from the training data in the dataset. Using unlabeled data, unsupervised learning creates models that may use knowledge and patterns that weren't previously understood. The dataset used to forecast the stage of lung cancer is an extensive compilation of information from 1000 people, each of whom is defined by 24 unique variables. These characteristics include a range of clinical, demographic, and diagnostic elements that are important for predicting the stages of lung cancer. The dataset has been carefully selected to facilitate the accurate classification of lung cancer into three stages: low, medium, and high. This classification is based on data from the last column, "Level." A thorough examination of the course and severity of lung cancer in the sampled population is made easier by the categorization of the disease into discrete phases. Many different types of information are contained in the dataset's attributes, some of which are patient-specific details like age, gender, smoking history, family medical history, vital statistics, results of medical tests, radiological

imaging findings, histological markers, and other clinical parameters. This dataset is an essential tool for carrying out in-depth analysis and creating prediction models using machine learning techniques. Through the utilization of this information, scholars and professionals may investigate associations, tendencies, and prognostic patterns to progress our comprehension of lung cancer staging and therefore improve early identification, treatment strategizing, and patient outcome. Because of the rich and varied nature of the dataset and its emphasis on lung cancer staging, it has great potential to aid in the development of precise, dependable predictive models that will improve patient outcomes and healthcare procedures by having a substantial impact on the early detection and management of lung cancer.

## **1.2 Motivation**

Lung cancer prediction models provide a viable way to address the difficulties related to early identification and treatment. These models seek to enhance patient outcomes, maximize healthcare resources, and support the overarching objective of lessening the burden of lung cancer on people and society by using cutting-edge technology and data analysis. Given that lung cancer affects most people, it is becoming more widespread and its prevalence rate is rising every day. Diet, exercise, and other aspects of lifestyle have a role in the development of lung cancer. In Bangladesh, lung cancer affects around 190,000 persons annually. It is anticipated that this malignancy would claim the lives of over 30,000 individuals. According to a 2020 study, out of 400 cancer patients, 11 had lung cancer in 2016. The male to female ratio was 10:1. Lung cancer is estimated to have killed 127,070 people (59,910 women and 67,160 men). Many researches have been conducted to forecast the incidence of lung cancer. I do a number of researches about lung illness prognosis. Most of them don't have any greater accuracy. I therefore became more resolute and eventually ascertained the highest level of precision in our process. I have created a method to predict whether lung cancer will manifest in people who are at risk or who are already receiving treatment.

### **1.3 The rationale of the study**

I am using pretend model in my study to forecast a patient's prognosis for lung cancer. I've just learned that this disease is starting to have an impact on our society. However, what I found was a lack of information and diagnostic resources. Both evaluating a patient's symptoms and making a lung cancer diagnosis are expensive procedures in our developing nation. As a researcher, my goal is to use machine learning to solve the issue.

### **1.4 Research Questions**

What's the operation of the algorithms in this suggested model?

What is the likelihood of survival for a person with lung cancer?

How can we predict lung cancer disease early detection?

What benefits does our suggested model provide?

What real-world applications are there for this work?

What is the anticipated future course of the project?

Which safety precautions are necessary in order to carry out this work?

How can my lung cancer prediction model's accuracy be evaluated?

To what extent is this work complicated?

What credentials are required for this position?

### **1.5 Expected Output**

The prevalence of lung cancer in the general population is rising. Moreover, nobody is certain whether she is impacted or not. Based on the diagnostic report, i provide the optimal approach for either detecting or forecasting the condition. My approach can accurately quantify the impact, enhance decision-making, and identify people with lung cancer. It may be used to measure associated challenges in addition to life quality. It could raise people's understanding of lung cancer in general. The suggested model has the fastest time to assess the illness.

## **1.6 Project Management and Finance**

My suggested technique offers a sensible and economical approach suitable for everyday use. Implementing this method for evaluating lung cancer cases could prove advantageous for our nation. While employing high-configured tools may yield optimal results and smoother functioning of the model, it remains feasible to utilize simpler tools. Regardless of the tool's sophistication, the model can still be effectively applied in real-life scenarios.

## **1.7 Report Layout**

Chapter 2 outlines the specifics of the previous researchers' study. Before delving into the investigation, I review the introduction and motivation sections. The introduction provides a detailed explanation of the proposed approach, while the motivation section offers insight into the rationale behind the prediction. Once the introductory part is complete, I focus on a thorough investigation, gathering internal data for my study. In the methodology section, I choose machine learning algorithms, apply them to my dataset, and determine the most effective ones. I carefully examine data during pre-processing stage and ultimately achieve the intended outcome, referred to as the comparison result, which is discussed in the conclusion section.

## CHAPTER 2

### BACKGROUND

#### 2.1. Preliminaries

Within this segment, I aim to identify the precise location of lung cancer using machine learning techniques. I delve into investigations concerning the evaluation and review of patients' diagnostic reports. Five distinct algorithms are employed for this purpose: Decision Trees, KNN, SVM, ANN, and LR. The study utilizes machine learning models to carry out the investigation, drawing inspiration from previous scholars who have employed multiple models in similar studies.

#### 2.2. Related works

A few artificial intelligence classifiers that I've used to diagnose lung disorders are appropriate for the study i am proposing. Chaturvedi et al. (2021b): Utilized LUNA16 and LIDC-IDRI datasets for detection, employing Gabor filtering and edge enhancement techniques. Achieved 80-90% accuracy using SVM, CNN, and ANN models. Kadir and Gleeson (2018): Worked on recognition using the 2017 Data Science Bowl (DSB) competition dataset by Kaggle. Employed Convolutional Neural Networks (CNNs) and achieved accuracy in the low 90s with AUC points around 0.85-0.87. Adetiba and Olugbara (2015): Employed SVM and ANN models on IGDB.NSCLC and COSMIC databases for detection, achieving a high prediction accuracy of 95.90%. Podolsky et al. (2016): Worked on classification using the Dana-Farber Cancer Institute Dataset. Achieved high accuracy ranging from 0.85 to 0.97 using KNN, SVM, Decision tree, and Naïve Bayes models. Patra (2020): Employed RBF on the Lung Cancer Dataset (UCI Machine Learning Repository) and achieved 81.25% accuracy, focusing on improving predictive accuracy. Morgado et al. (2021): Focused on prediction using the NSCLC-Radio genomics Dataset, utilizing PCA methods and SVM, Logistic Regression models with accuracy ranging from 0.725 to 0.737. Yakar et al. (2021): Predictive study using a dataset from Radiation Oncology and Chest Diseases, achieving 92% accuracy with the LGBM model for predicting RP. Gould



et al. (2021): Explored early detection using data from 6,505 NSCLC case patients, obtaining 95% accuracy in identifying high-risk individuals through machine learning models. Δρίτσας and Trigka (2022): Focused on early detection and achieved high accuracy rates of 97.1% to 99.3% using SMOTE, Rotation Forest, KNN, and ANN models. Benzekry et al. (2021): Employed logistic regression, random forest, single-layer neural network, naive Bayes, and KNN models on patient data, aiming to enhance predictive performance.

### 2.3. Comparative Analysis and Summary

At the moment, a lot of people employ the machine learning model. I had to start the difficult task of finding our relative's place of employment. Every related effort created poor-quality models with low accuracy. Applying many Machine Learning models was necessary to get the highest level of accuracy in predicting the dataset. I need to employ high-configuration equipment to run the models. I calculated the classification rates using a number of different methods. Adding costly GPUs to complex models may result in long runtimes.

Paper Name (Year)	Working models (My work)	Working models (Other works)	Accuracy (My work)	Accuracy (Other works)
Chaturvedi et al. (2021b)	Logistic Regression, Decision Tree, KNN, ANN, SVM	SVM, CNN, ANN	100%	80-90%
Kadir and Gleeson (2018)	-	CNNs	-	low 90s AUC points
Adetiba and Olugbara (2015)	SVM, ANN	SVM, ANN	97.50% 99.50%	95.90%

Patra (2020)	-	RBF	-	81.25%
Morgado et al. (2021)	SVM, Logistic Regression	SVM, Elastic Net, Logistic Regression	97.50% 100%	72.5-73.7%
Yakar et al. (2021)		LGBM		92%

Table 3.2: Comparative Analysis

#### 2.4. Scope of the Problem

The issue was making the procedure of diagnosing lung cancer sickness more straightforward and familiar. Owing to the abundance of research using machine learning, I try to deliver the best accuracy I can with my recommended model. Even if there wasn't much space for improvement, the concept might be used to lower the frequency of lung cancer diagnosis by using some basic technologies.

#### 2.5 Challenges

On the GitHub website, the collection was housed. That was really easy to use and the information was helpful. After gathering all the data, I have to manually check the dataset to make sure nothing is missing. My accuracy with this dataset is unmatched by anybody else.

## CHAPTER 3

### RESEARCH METHODOLOGY

#### 3.1 Research Subject and Instrument

To maximize accuracy from the dataset, I employed a variety of hybrid models and techniques. Efficient setup tools with top-tier GPUs were essential for this endeavor. I utilized programming tools like Google Collaboratory and the Python programming language. These platforms enable the creation and execution of Python code directly in a web browser, offering unparalleled flexibility and convenience.

#### 3.2 Data Collection Procedure

The dataset was gathered via GitHub, it was almost ready for usage. This table has 1000 rows and 24 columns. The diagnostic column is used to categories lung cancer incidence. Each characteristic aided in the identification and assessment of lung cancer. Three classes of patients are identified: low, medium, and high. Therefore, Low signifies a low level of lung cancer in the patient, Medium denotes a medium level, and Third denotes a high degree of lung cancer in the patient. I put these three situations' frequencies at low, medium, and high. Data of 1000 patients is given in this dataset and 21 attributes ranged from 1-9 scale. Of these, 1-3 low, 4-6 medium and 7-9 high people are rated on a scale. Based on these measurement's, the patient's level is overall determined as low, medium, high. The ratio is shown in figure 3.1 that follows. They've undergone testing and training. I made the decision to devote 20% to testing and 80% to training.

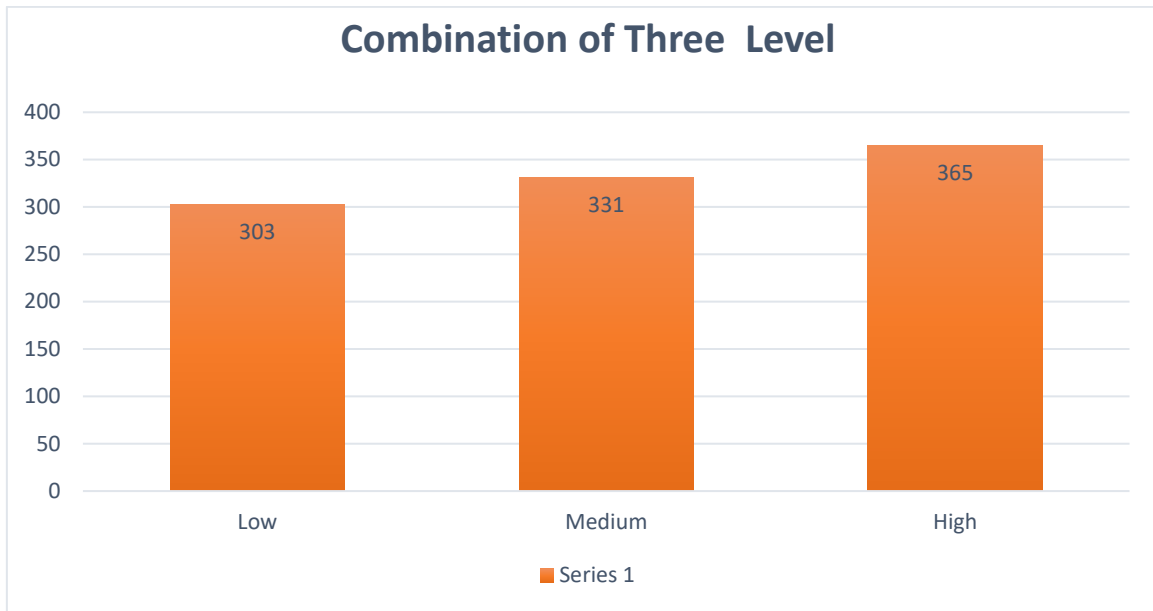


Figure 3.1: Number of target value

Here is the combination of three level of lung cancer patient's dataset. In my dataset there are 303 low level patients and 331 medium level patients and 365 high level patients.

The dataset included nominal values and neither erroneous nor missing values.

Table 3.1: Details of the dataset

Column	Non-Null Count	Dtype
Patient Id	1000 non-null	object
Age	1000 non-null	int64
Gender	1000 non-null	int64
Air Pollution	1000 non-null	int64
Alcohol use	1000 non-null	int64
Dust Allergy	1000 non-null	int64

Occupational Hazards	1000 non-null	int64
Genetic Risk	1000 non-null	int64
Chronic Lung Disease	1000 non-null	int64
Balanced Diet	1000 non-null	int64
Obesity	1000 non-null	int64
Smoking	1000 non-null	int64
Passive Smoker	1000 non-null	int64
Chest Pain	1000 non-null	int64
Coughing of Blood	1000 non-null	int64
Fatigue	1000 non-null	int64
Weight Loss	1000 non-null	int64
Shortness of Breath	1000 non-null	int64
Wheezing	1000 non-null	int64
Swallowing Difficulty	1000 non-null	int64
Clubbing of Finger Nails	1000 non-null	int64
Frequent Cold	1000 non-null	int64
Dry Cough	1000 non-null	int64

### **3.2.1 Categorical Data Encoding**

Nominal values are created from categorical data using the categorical encoding scheme. I used solely numerical data for input and output in our machine learning study, therefore the categorical encoding approach was crucial. I can utilise all of the columns for the categorical data encryption strategy, with the exception of AGE.

### **3.2.2 Missing Value Imputation**

It comprises imputed numbers to fill in any gaps or missing data found via examination of the other dataset. However, my data's lack of null values is comforting.

### **3.2.3 Handling Imbalanced Data**

The procedure for altering a data source's category distribution. By gradually adding additional instances, it controls the dataset. When the whole dataset is used the amount of data for minorities is raised as input.

### **3.2.4 Feature Scaling**

It is a technique for bringing several different independent data sets into one normalized state. The data level measure scales all relevant data in low, medium, high.

## **3.3 Statistical Analysis**

A research project's analytical component is essential. This part depends on the development and assessment of the algorithms I've employed. Since I want to use a comma-separated values (CSV) file, I must first clean and prepare the data set for usage. I used a range of procedures, such as gathering data and pre-processing, among others. In this paper, I used five different kinds of algorithms: Decision Trees, K-Neighbors Classifiers (KNN), Support Vector Machines (SVM), Logistic Regression (LR), and Artificial Neural Networks (ANN). I also used a range of ensemble models to support the performances. The two models with the greatest accuracy, LR and Decision Tree, were 100%. Next came KNN, ANN, and SVM, which had a 99.5%, 99.5%, and 97.5% accuracy. Ten-fold cross-validation and hyperparameter adjustment were used.

### 3.4 Proposed Methodology Flow chart:

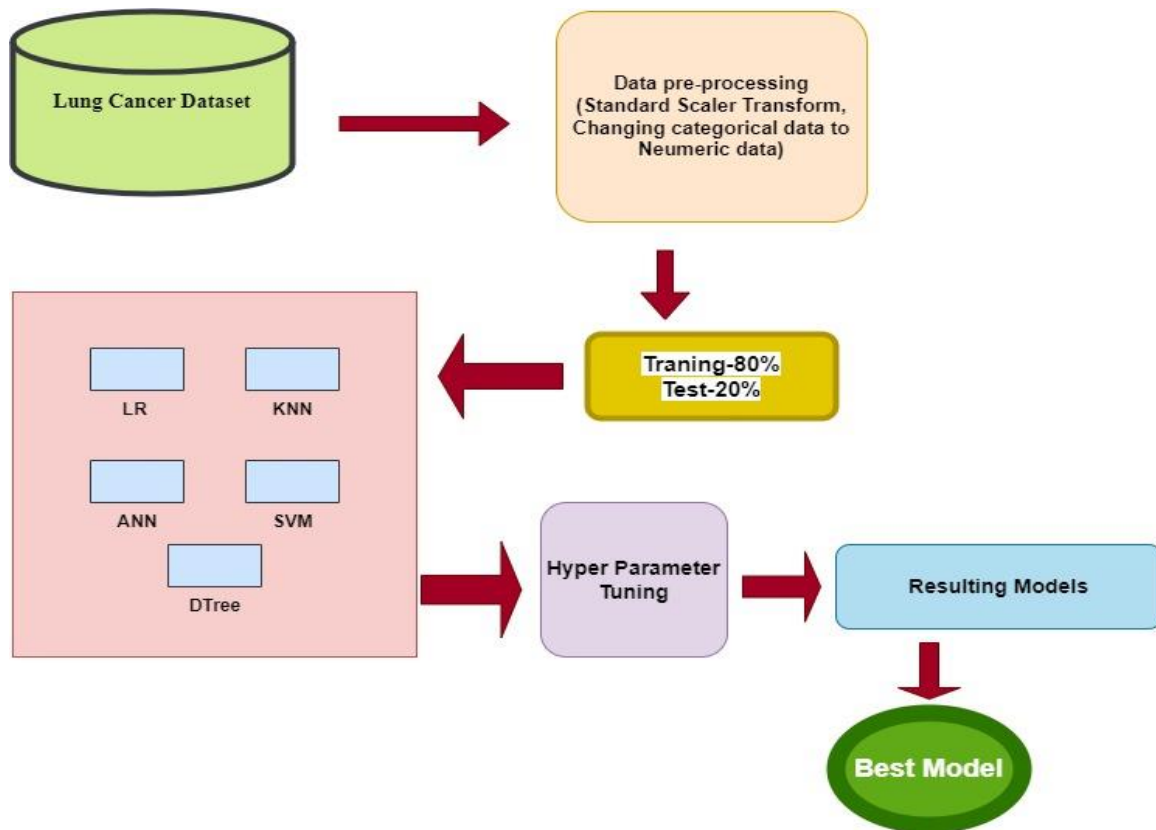


Figure 3.2: Methodology of Lung Cancer disease

Using a process diagram, I made a prediction about lung cancer sickness in this part. In order for the system to be assessed and trained, I started by inputting the dataset. The data was then prepared using the Standard Scaler Transform. transforming category information into a numerical format. Of these, 20% were utilized for testing and the remaining 80% for instruction. I then used algorithms to assess the outcomes. I then utilized ensemble approaches to get the maximum forecast accuracy. Subsequently, the outcomes of the used ensemble methods were assessed. After that, the results were verified via hyper parameter tuning. Following that, I assessed the models that had been used and made decisions based on the outcomes. Figure 3.2 depicts our technique's whole operation.

A correlation subplot may reveal intrinsic dependencies between variables that alter their relationship to one another. The more connected the variables are, the more probable it is

that one may be predicted from the other. It enhances our ability to recognize significant information and points to a better comprehension of the dataset.

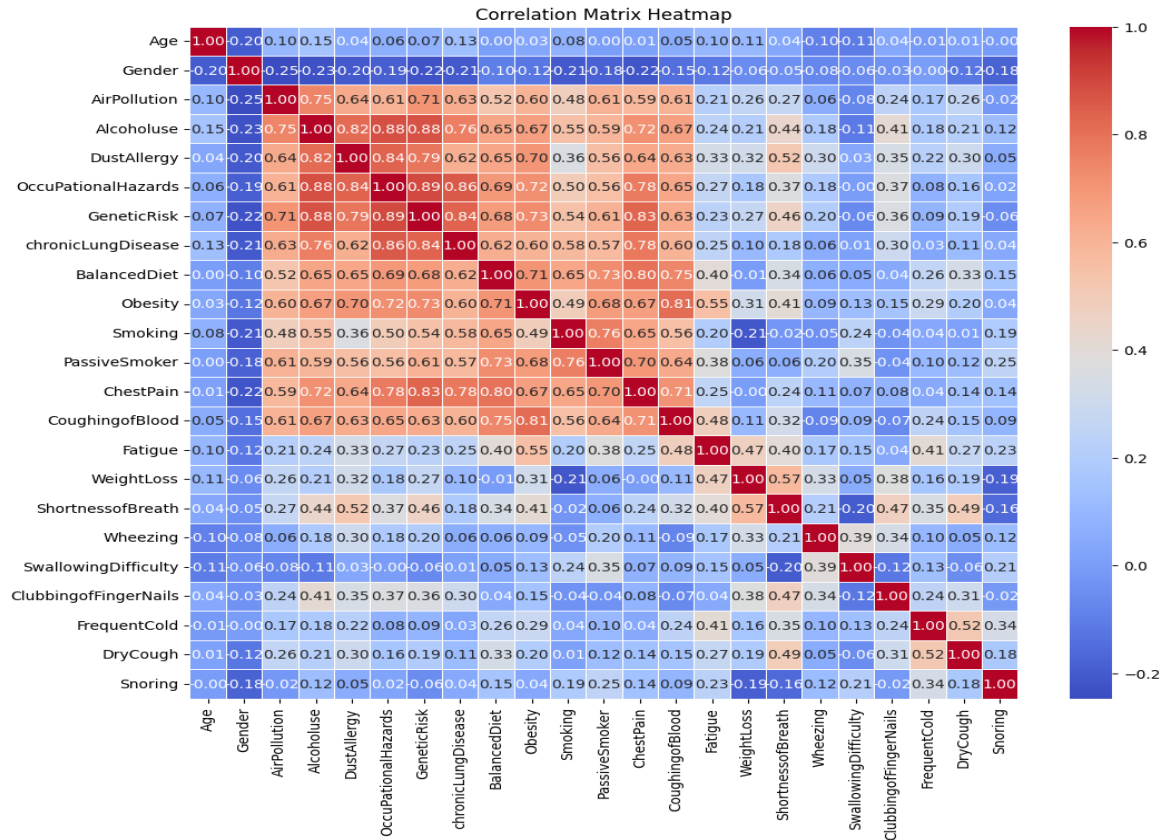


Figure 3.3: Correlated Features of Lung cancer disease Dataset

### 3.5 Implementation Requirements

In order to utilize our proposed model effectively, I require reliable data sources for study or model training. Cleaning the dataset is imperative to ensure proper functionality. Various filtering techniques are employed for dataset cleaning. One of the subsequent methods utilized for data preprocessing is the Standard Scaler Transform, which involves converting numerical data from categorical data. Eighty percent of the data is allocated to training, while the remaining twenty percent is reserved for testing purposes. Following this, algorithms are implemented and their outcomes assessed. Ensemble techniques are then employed to enhance prediction accuracy. The outcomes of the algorithmic ensemble are carefully evaluated. Hyperparameter tuning is subsequently employed to validate the results. Finally, an assessment of the models implemented using the data is conducted.



## CHAPTER 4

### EXPERIMENTAL RESULTS AND DISCUSSION

#### 4.1 Experimental Setup

This study used a training-and testing-based supervised learning approach. The classification model is constructed with the aid of the training dataset. To get the outcome, the testing dataset is subjected to the constructed model. The next parts will provide a brief illustration of the machine-learning technique.

##### 4.1.1 Classifier Algorithms

I have used different five types of algorithms like LR, Support Vector Machine, K-Neighbors (KNN), ANN and Decision Tree.

##### **Support Vector Machine (SVM)**

Support Vector Machine (SVM) is a supervised machine learning technique commonly used for regression and classification problems. SVM works by creating a hyperplane, or a set of hyperplanes, in a high-dimensional space to effectively separate different classes during classification tasks. In simple terms, SVM aims to find the optimal decision boundary that maximally separates data points belonging to different classes. This decision boundary is chosen to provide robust generalization to new, unseen data by maximizing the margin. SVM can handle both linear and non-linear decision boundaries by utilizing different kernel functions. The figure below illustrates how a decision boundary or hyperplane is utilized to classify two distinct groups:

The Support Vector Machine (SVM) operation is shown in Figure 4.1.

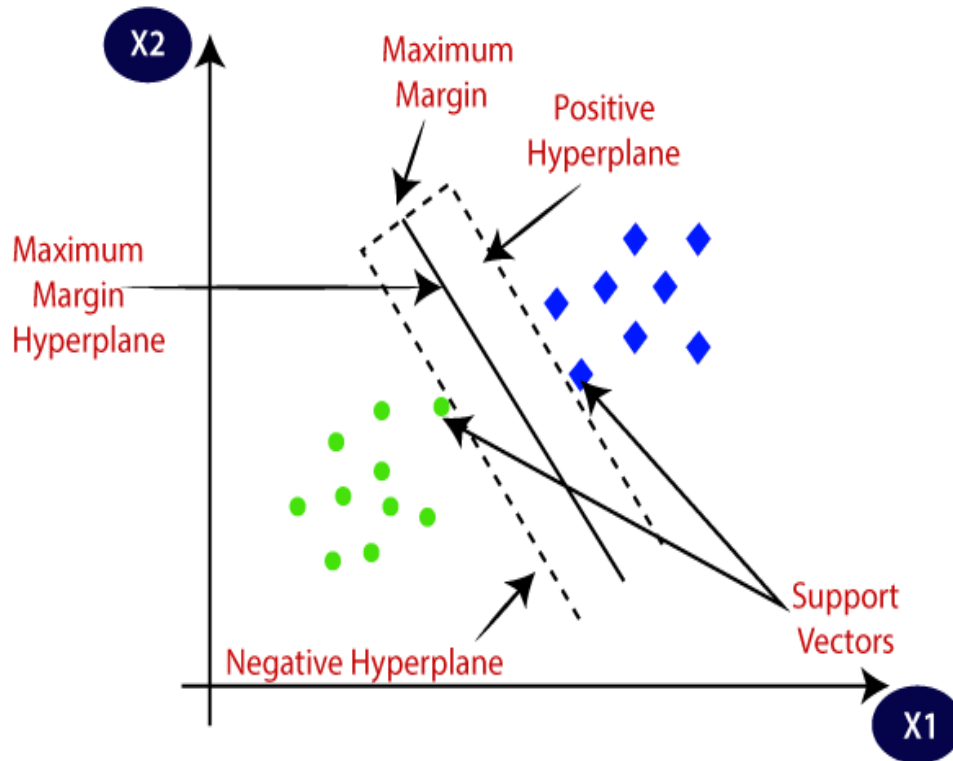


Figure 4.1: SVM classifier

### Artificial Neural Network (ANN)

ANN often referred to as ANNs, are computer models inspired by the structure and functioning of biological neural networks, particularly those found in the human brain. An artificial neural network comprises interconnected nodes organized into layers, also known as artificial neurons or perceptron's.

Weights are assigned to the connections between neurons, and each neuron applies an activation function to the total weighted sum of its inputs. In the training phase, the network gains knowledge by modifying these weights in response to errors or discrepancies between expected and observed results. Over time, the network may improve its performance and provide predictions that are more accurate thanks to a process called backpropagation.

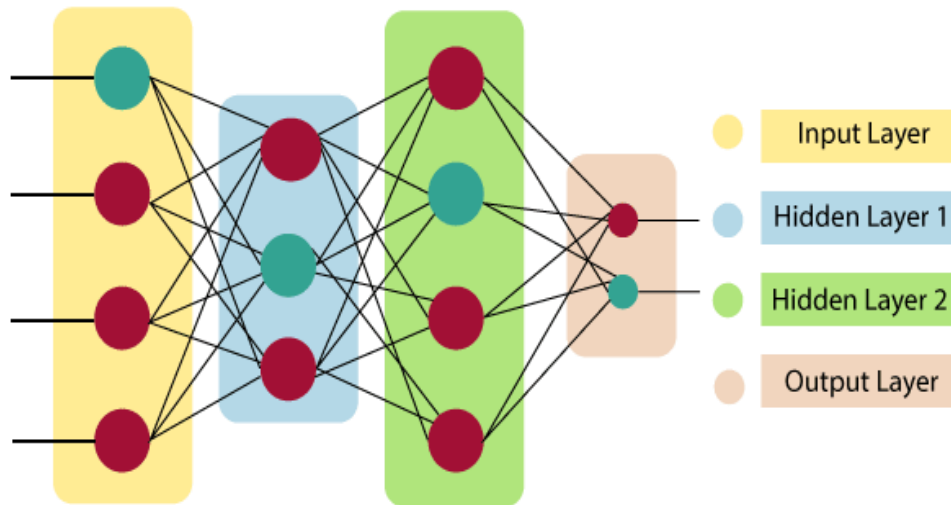


Figure 4.2: ANN Classifier

### Decision Tree

A supervised machine learning approach for classification and regression problems is called a decision tree. It creates a tree-like structure of choices by recursively dividing the data into subsets according to the most important characteristic at each node. Every internal node denotes a choice made in light of a certain property, and every leaf node shows the expected result. The objective is to use the input data to learn decision rules and build a tree that effectively classifies or predicts the target variable. Decision trees are useful for comprehending decision-making processes within a model because of their simplicity and interpretability.

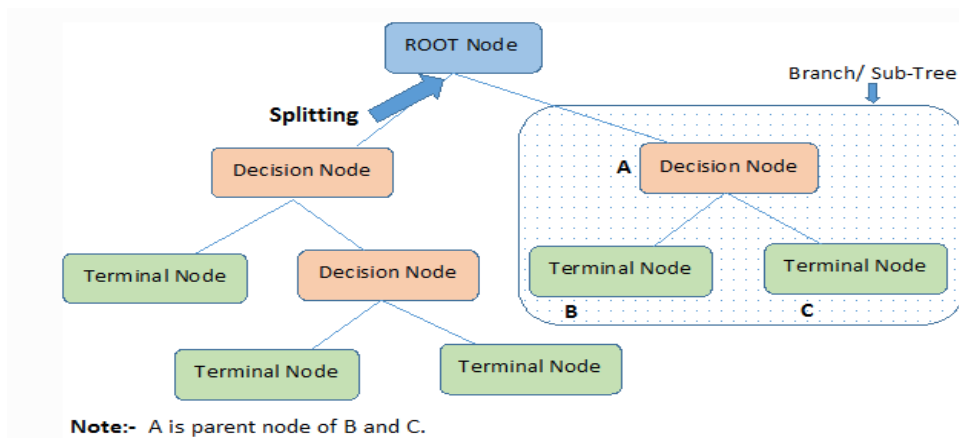


Figure 4.3: Decision Tree

## K-Nearest

K-Nearest Neighbors (KNN) is a machine learning approach that enables non-parametric classification algorithms to treat both new and old data equally. This concept is illustrated in Figure 4.4 below. KNN calculates the Euclidean distance between newly created data points  $(x_1, x_2)$  and existing data points  $(y_1, y_2)$ .

$$Distance = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}. \quad (5)$$

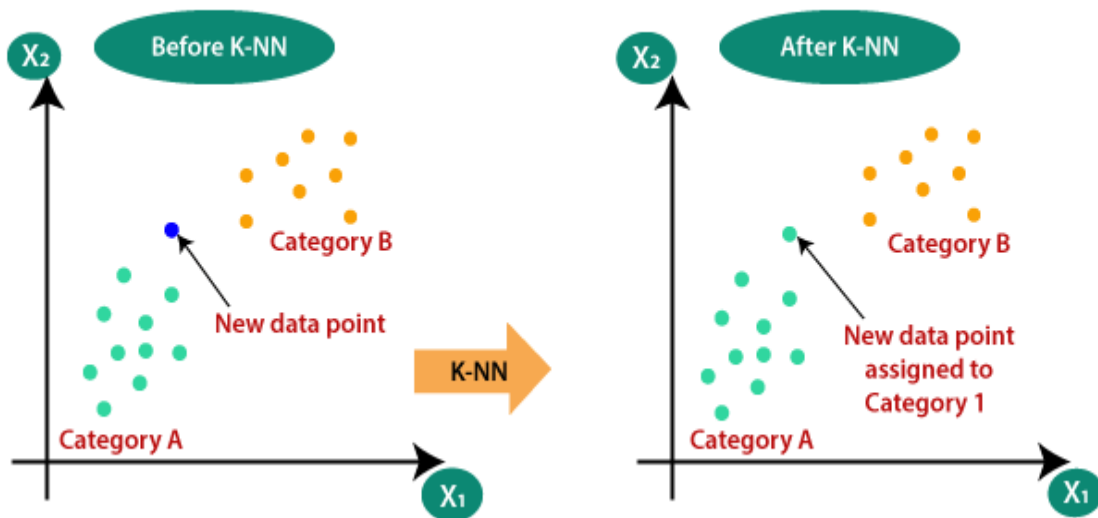


Figure 4.4: K-Nearest Classifier

## Logistic Regression

In logistic regression, the algorithm calculates the weighted sum of input features, applies a sigmoid activation function to transform the result into a probability, and then assigns the instance to the class with a probability greater than a specified threshold (often 0.5).

$$\text{The formula } h\theta(x) = -(\beta_0 + \beta_1 X) \quad (6)$$

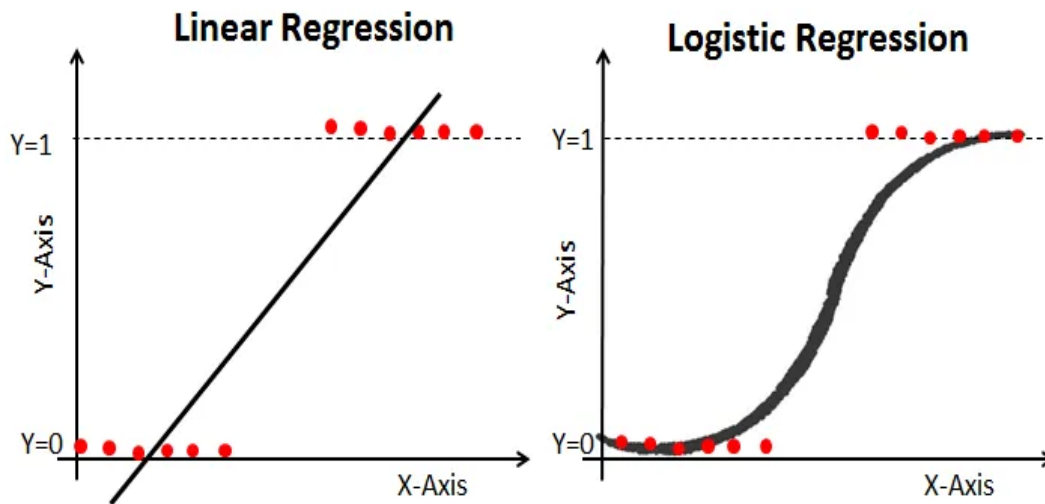


Figure 4.5: Logistic Regression Classifier

#### 4.1.2 Ensemble Methods of Machine Learning

A "ensemble approach" is the use of several classifiers to get the best accuracy and efficiency; this may turn weak algorithms into powerful classifiers. It was used in my investigation because of the variables pertaining to bias and uncertainty. It lowers the prediction spread, lowers variances, and aggregates predictions from several models. In my investigation, the four-ensemble approach was used.

#### 4.2 Experimental Result & Analysis

I had to evaluate the effectiveness of the available models at this stage. To make sure that my recommended model operates well, I may make use of a number of performance evaluation instruments. These algorithms evaluate the overall effectiveness using data that was not known before. I have to provide an analytical report in this part that is based on the experimental results of my machine learning models using the dataset i have chosen to study lung cancer sickness. I started by using the dataset i had selected. I computed the missing or erroneous values and removed them from my dataset. I tested a number of algorithms and evaluated their performance. For my suggested approaches, I assessed the Accuracy, Precision, Recall, and F1 Score of the Confusion Matrixes. These confusion matrices have been assessed for traditional techniques. I used five different kinds of

methods: Decision Tree, KNN, Artificial Neural Network, Support Vector Machine, and LR. I have seen several ensemble techniques in action, and I have also employed confusion matrices.

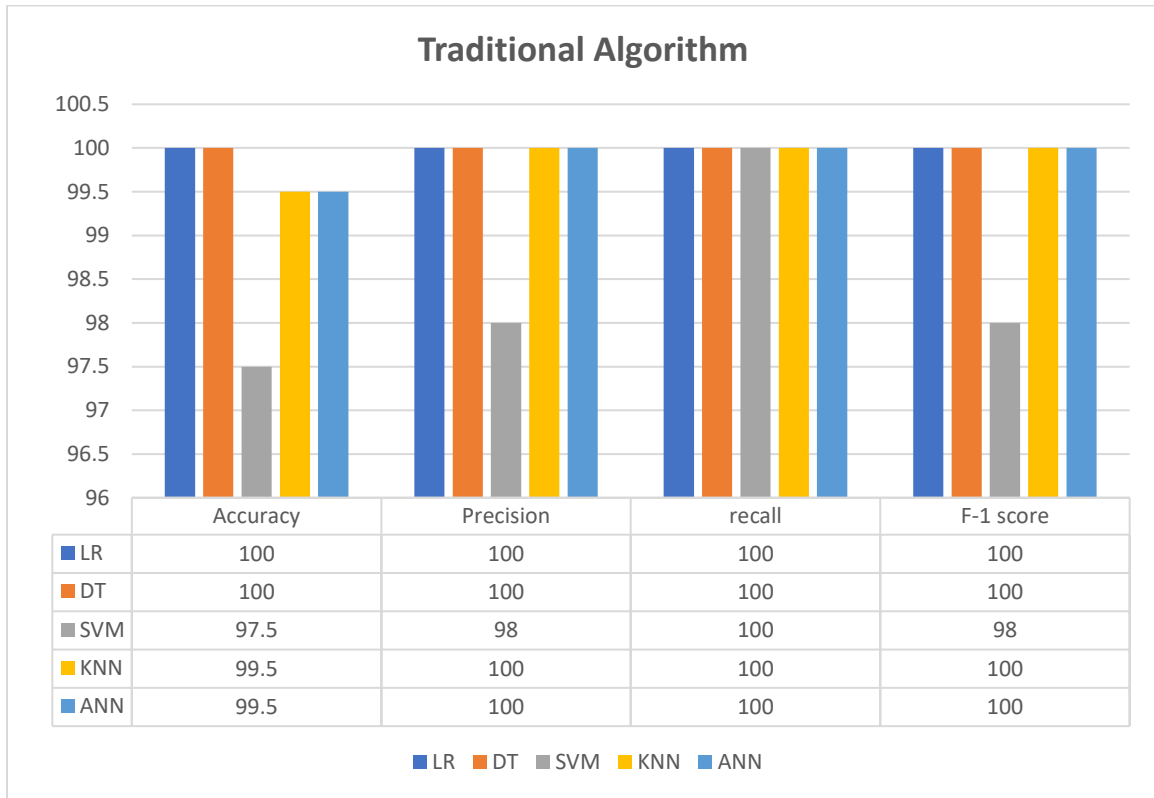


Figure 4.6: Experimental Results of Classifiers

The greatest accuracy was achieved with LR and DT, which had a 100% accuracy rate. SVM, KNN and ANN had accuracy rates of 97.50%, 99.50% and 99%, respectively. LR and DT provided the highest precision (100%), whereas SVM, KNN and ANN received precision values of .98%, 100%, 100%, and 100%, respectively. LR and DT provided the highest recall score (100%), while SVM, KNN and ANN had recalls of .98%, .100% and 100%, respectively. LR and DT provided the highest F-1 score (100%), while SVM, KNN and ANN received accuracy values of 98%, 100% and 100%, respectively.

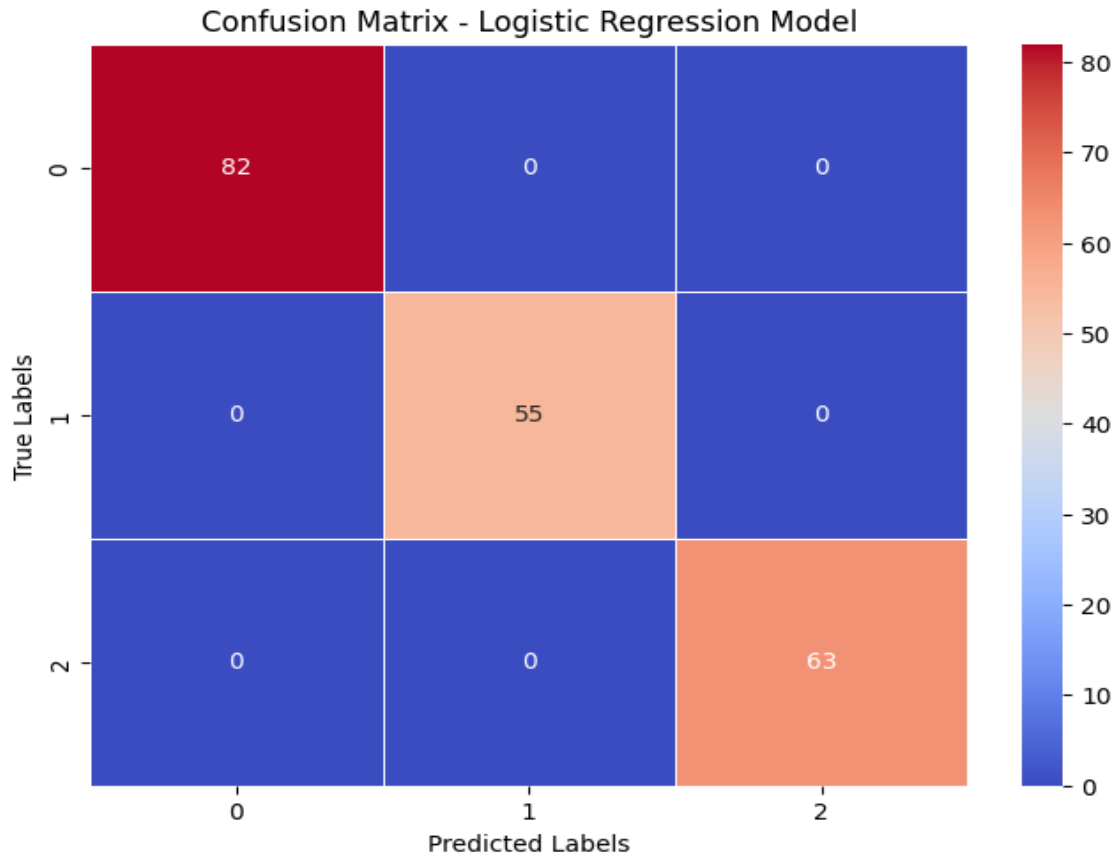


Figure 4.7: Confusion Matrix Analysis of Logistic Regression

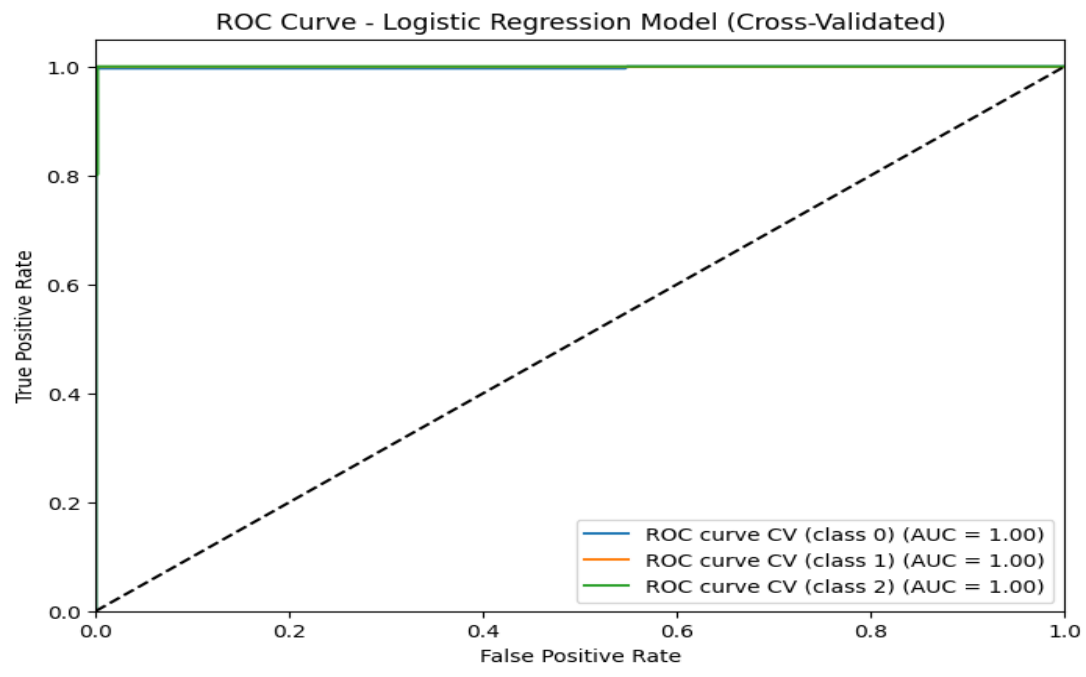


Figure 4.8: Cross Validation Analysis of Logistic Regression

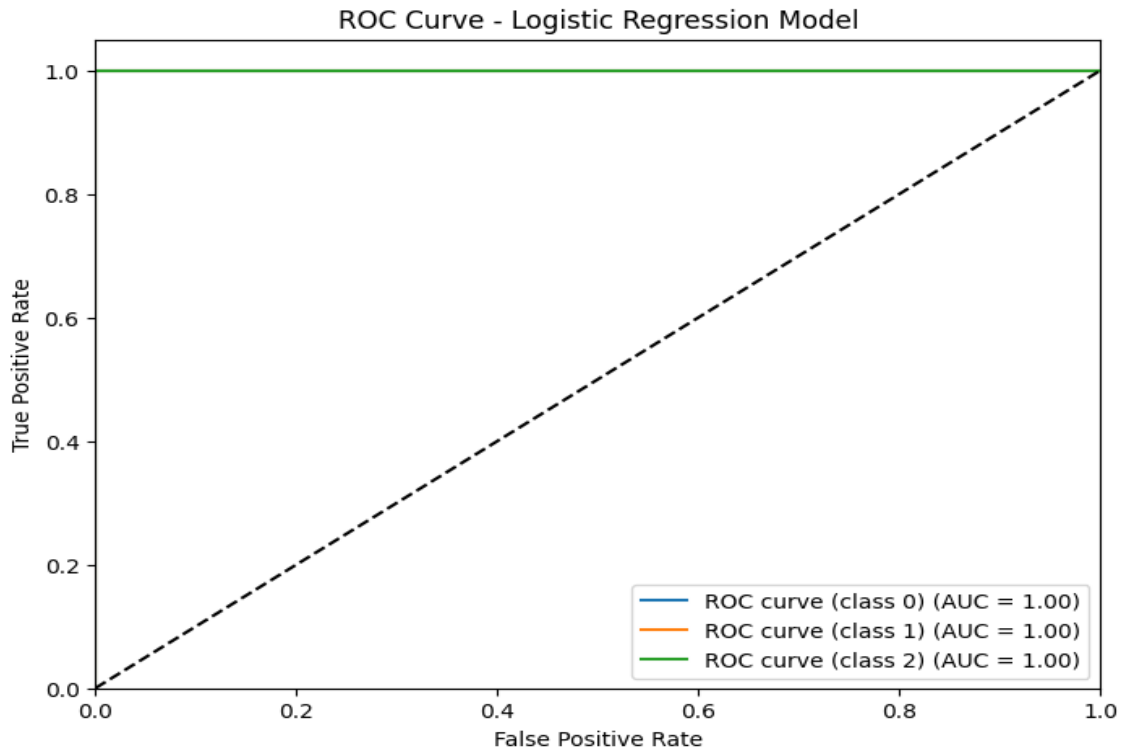


Figure 4.9: AUC-ROC Curve Analysis of Logistic Regression

First, I have applied a Logistic Regression algorithm for better calculation. Logistic Regression achieved the highest score, 100%. Precision had 100%, recall had 100% and the F-1 score 100%. which was a better score than SVM, KNN, ANN models. But the Decision tree model performed as well as these models.



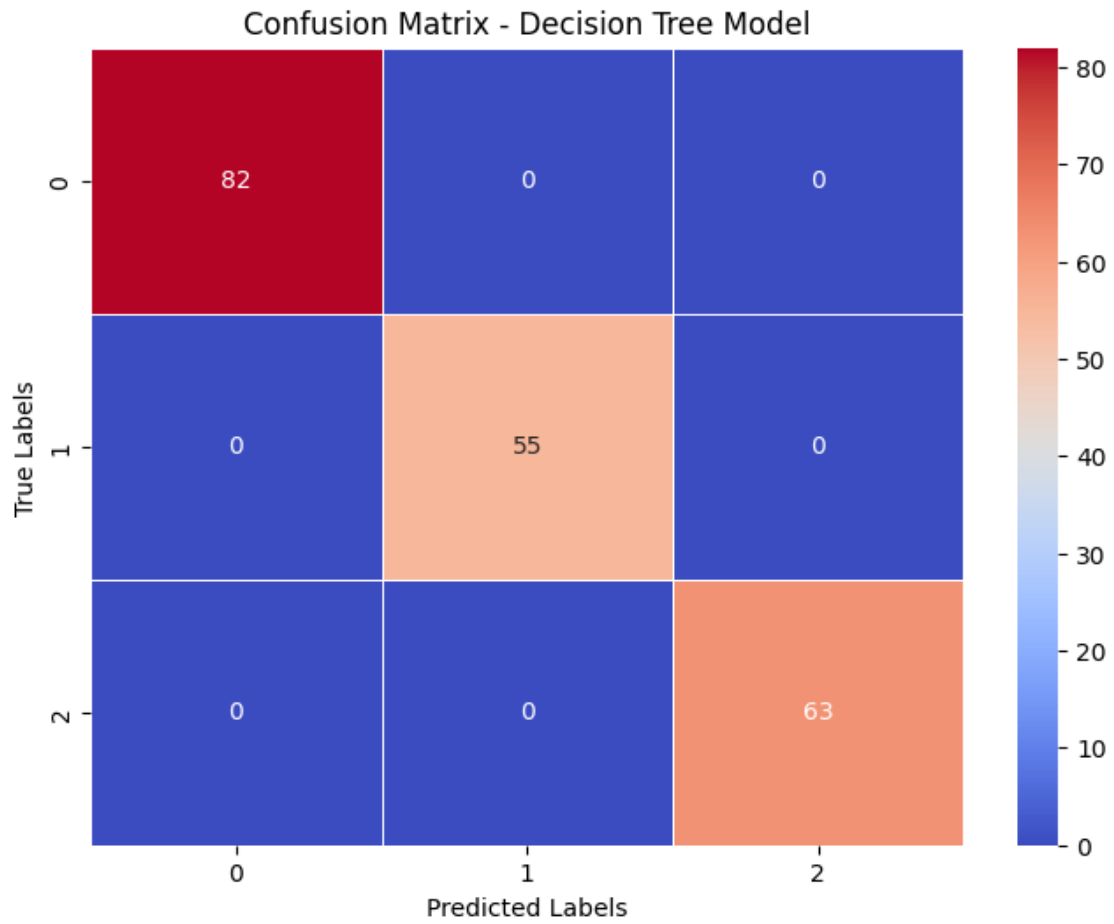


Figure 4.10: Confusion Matrix Analysis of Decision Tree

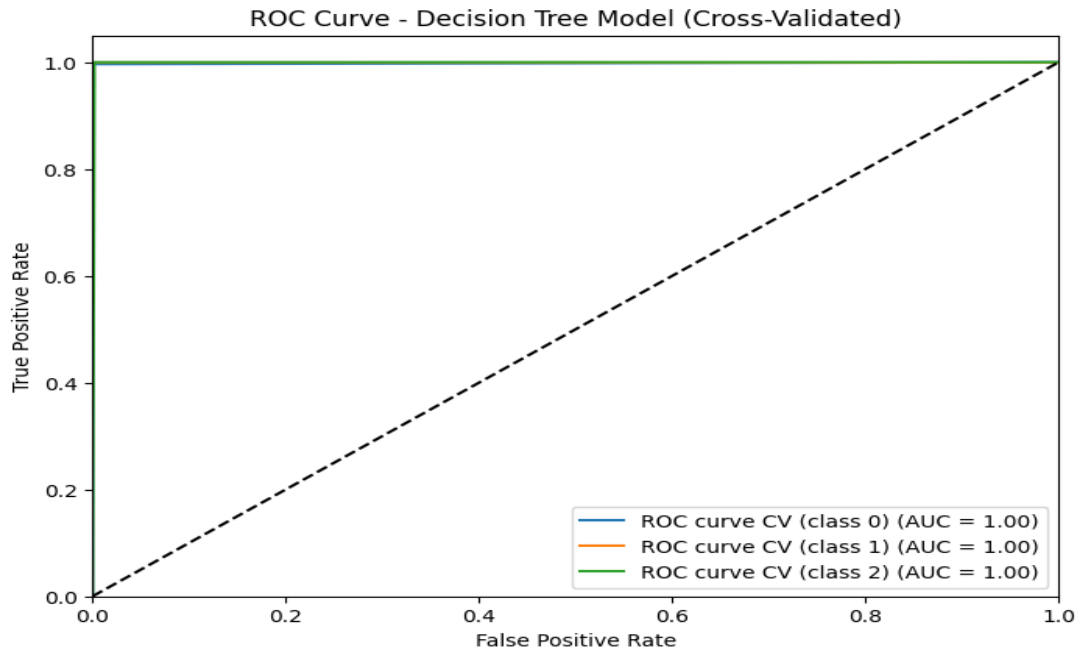


Figure 4.11: Cross Validation Analysis of Decision Tree

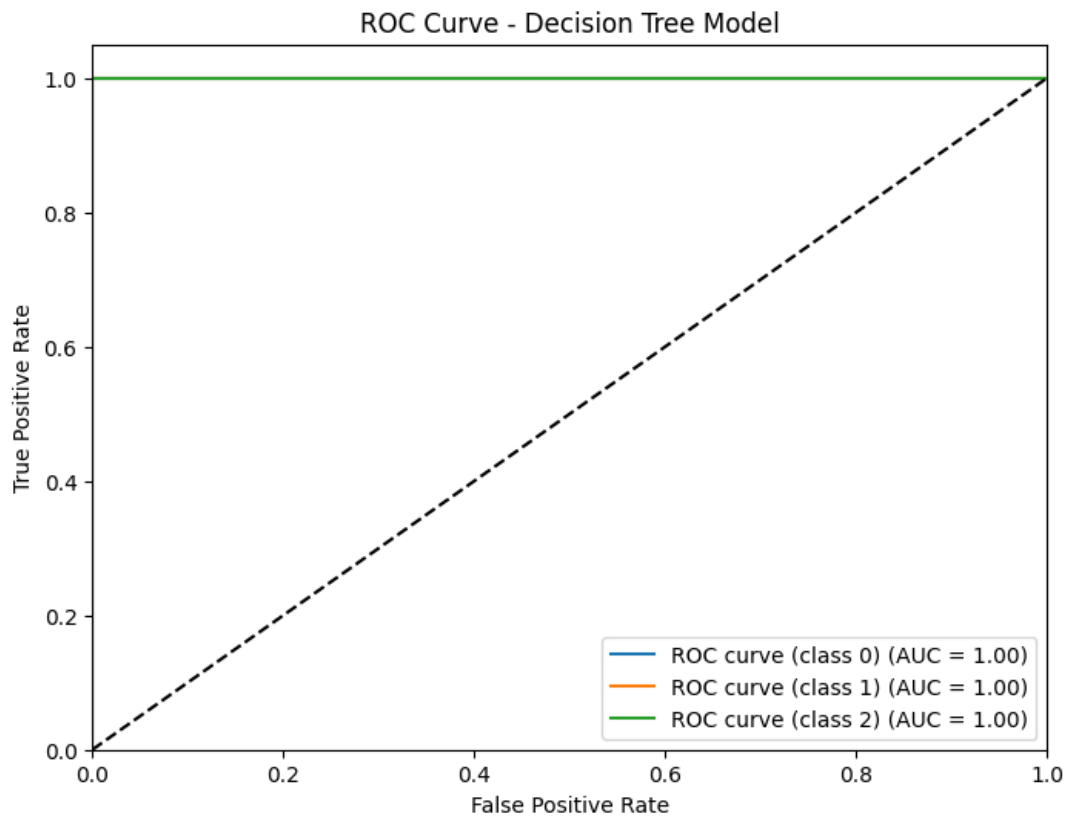


Figure 4.12: AUC-ROC Curve Analysis of Decision Tree

Second, I have applied a Decision Tree algorithm for better calculation. Decision Tree achieved the highest score, 100%. Precision had 100%, recall had 100% and the F-1 score 100%. which was a better score than SVM, KNN, ANN models. But these models performed as well as the same logistic regression model.

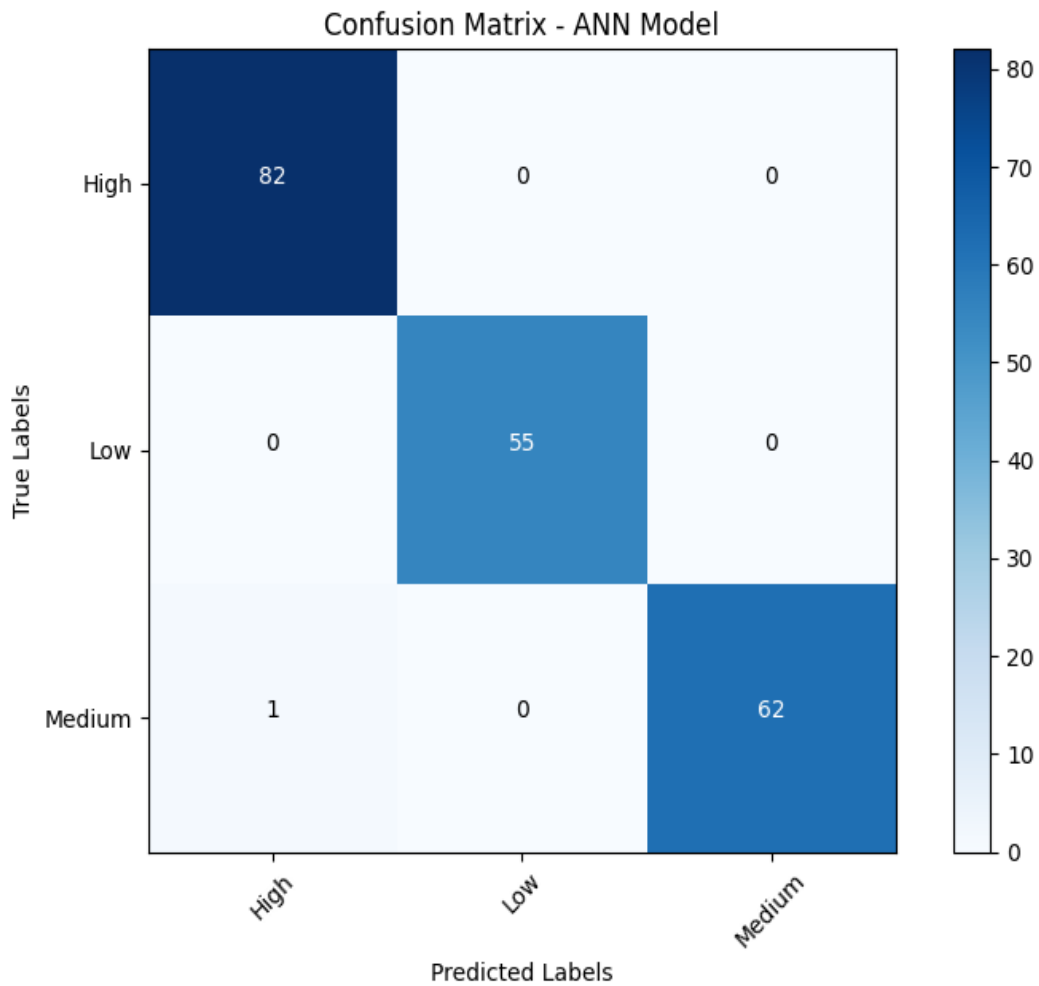


Figure 4.13: Confusion Matrix Analysis of ANN

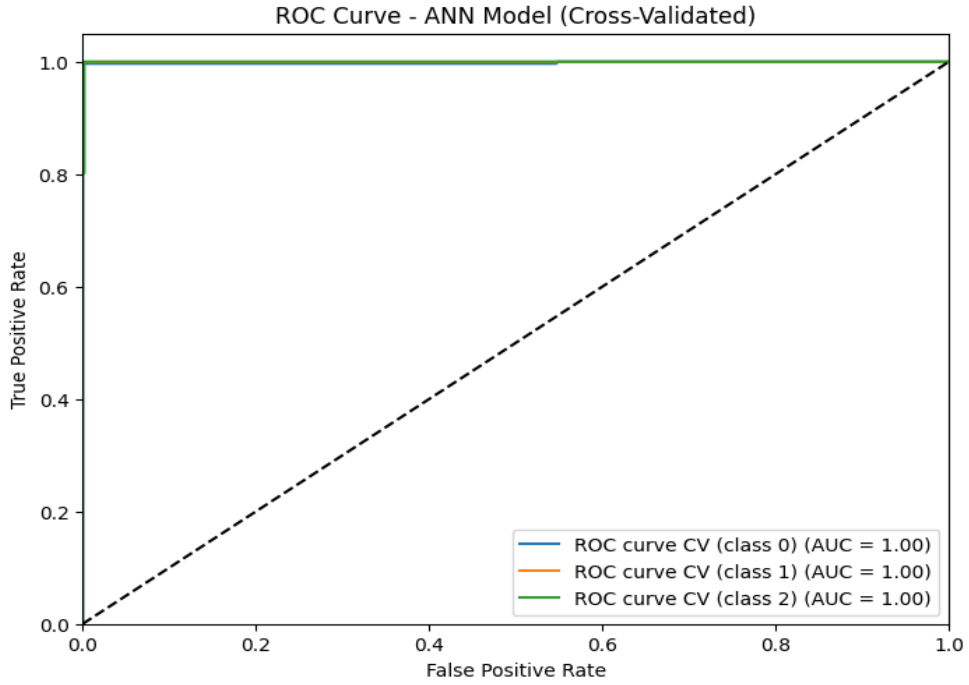


Figure 4.14: Cross Validation Analysis of ANN

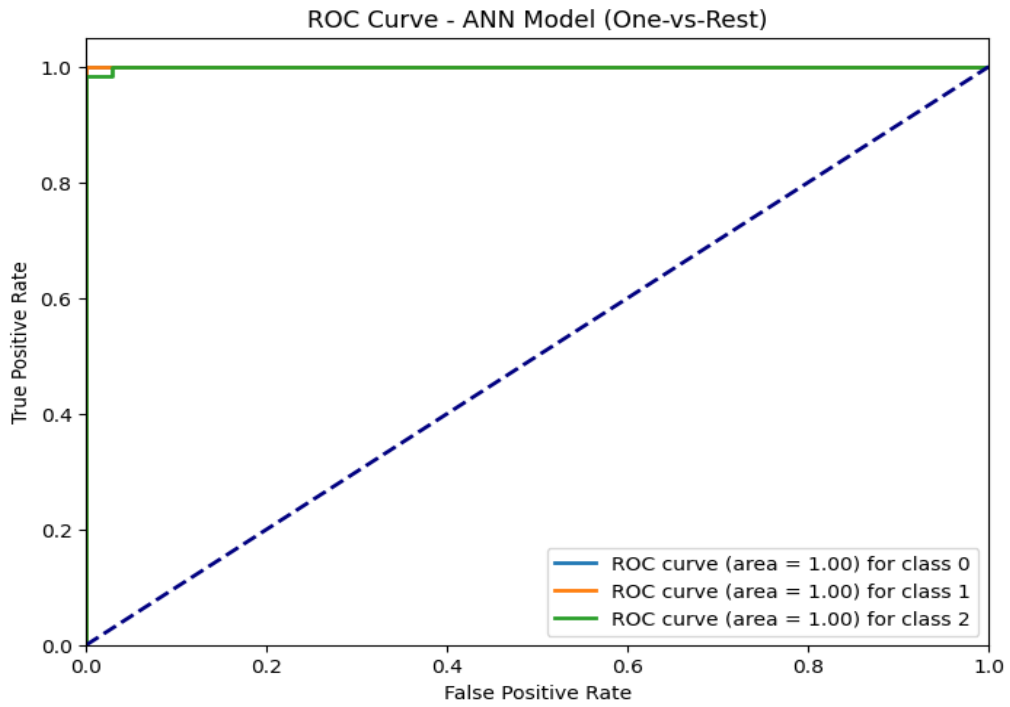


Figure 4.15: AUC-ROC Curve Analysis of Artificial Neural Network

ANN achieved the score, 99.50%. the precision 100%, recall had 100% and F-1 score had 100%. which was a better score than SVM models.

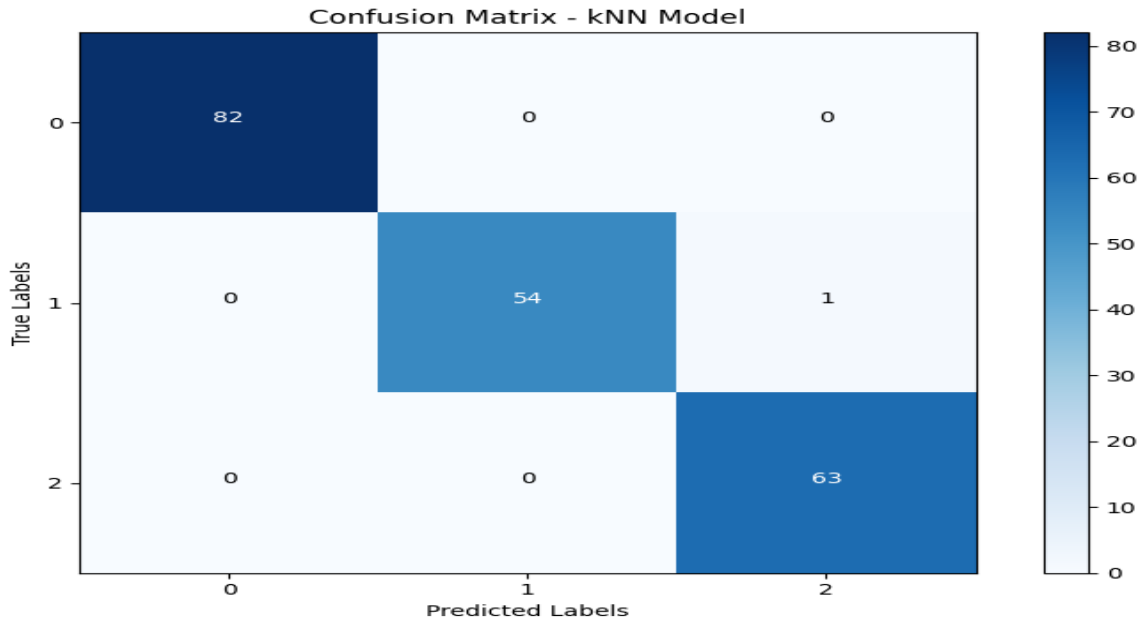


Figure 4.16: Confusion Matrix Analysis of KNN

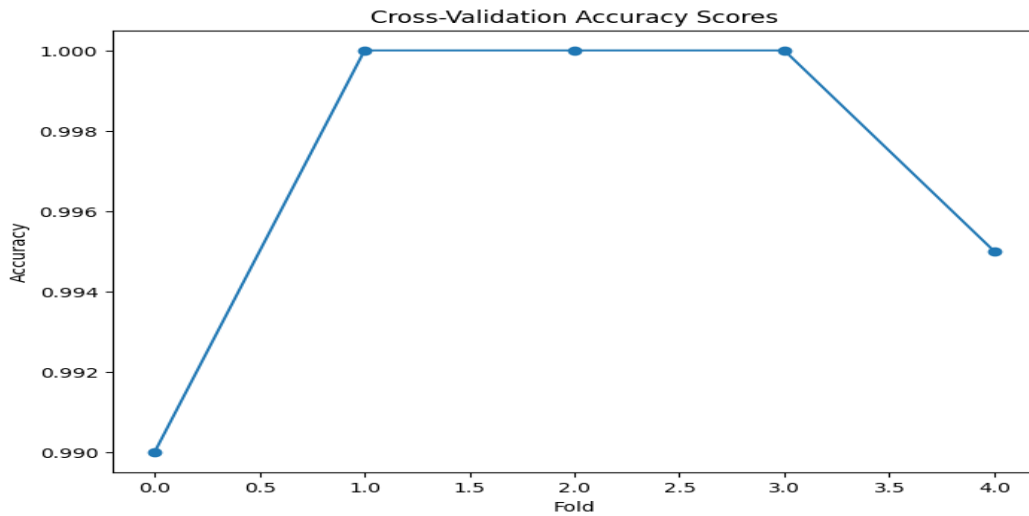


Figure 4.17: Cross Validation Analysis of KNN

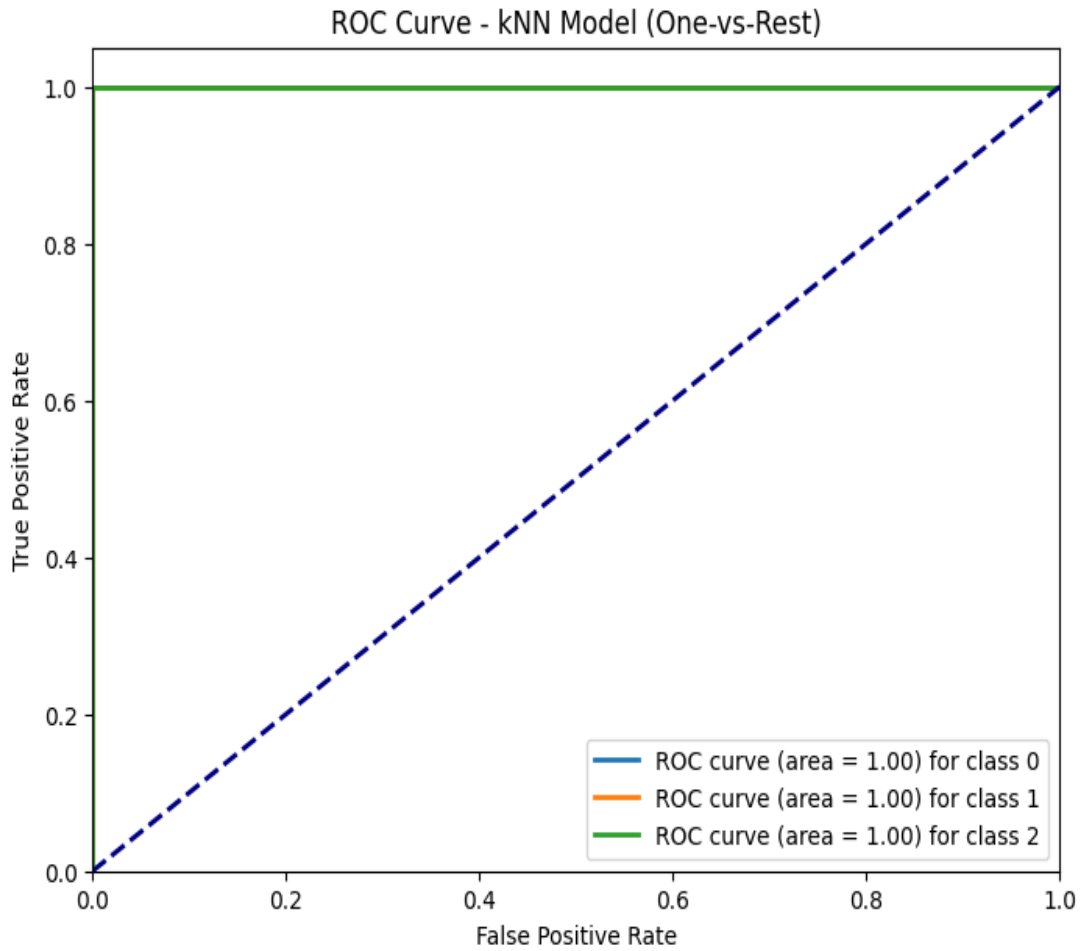


Figure 4.18: AUC-ROC Curve Analysis of KNN

KNN achieved the score, 99.50%. precision had 99%, recall 100% and the F-1 score had 99%. which was a same score by ANN models.

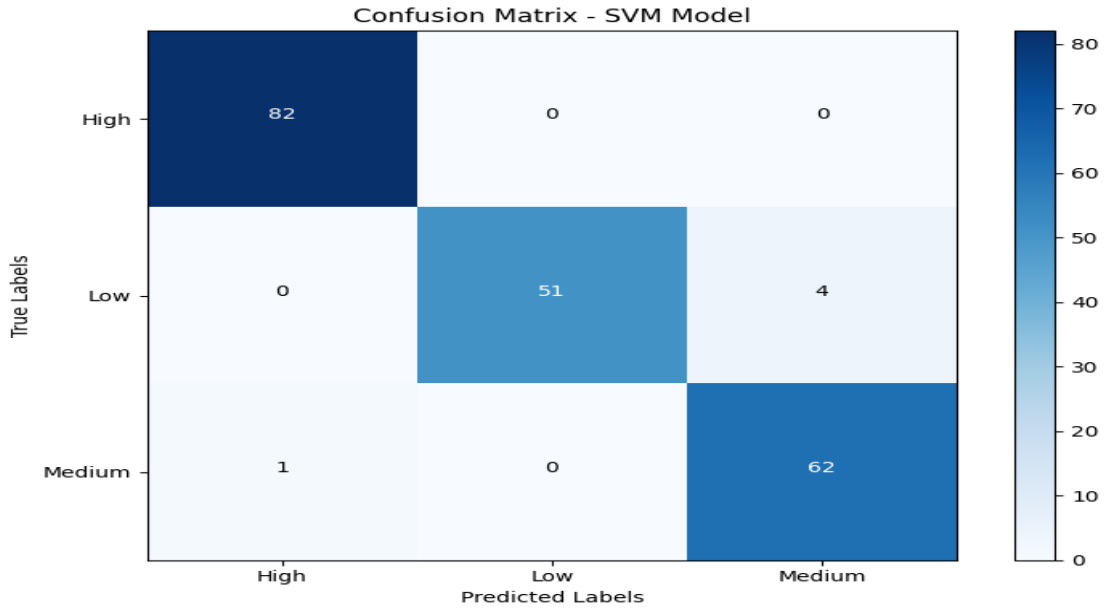


Figure 4.19: Confusion Matrix Analysis of SVM

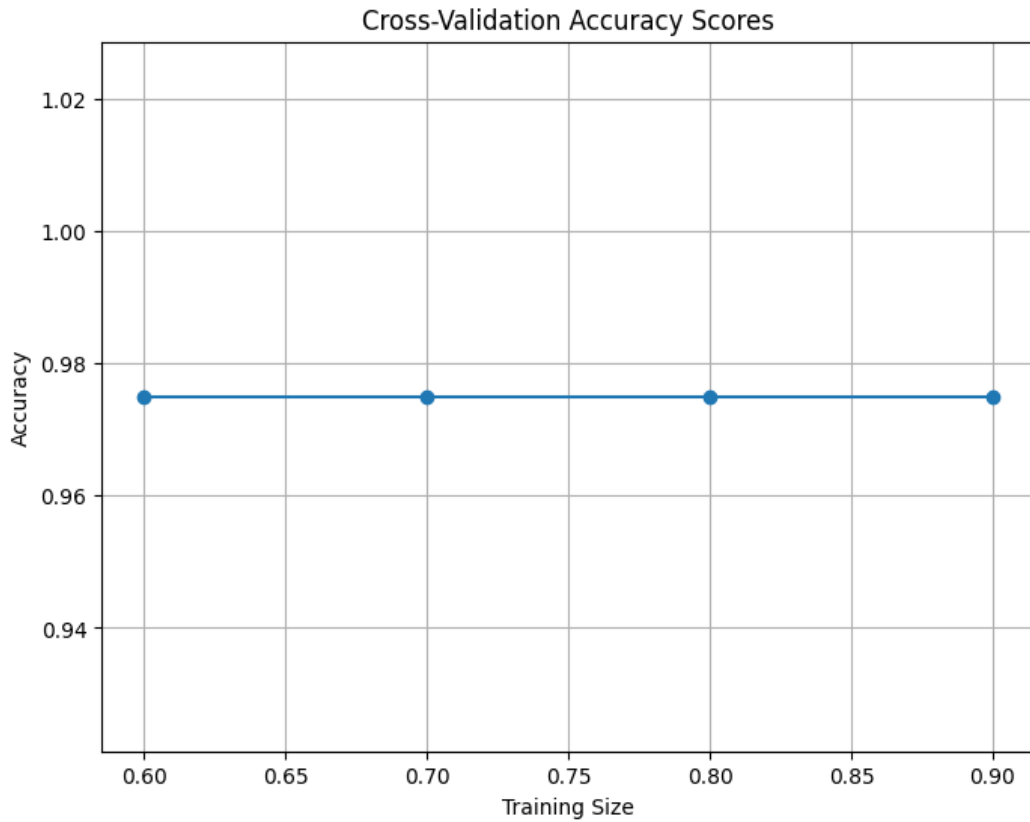


Figure 4.20: Cross Validation Analysis of SVM

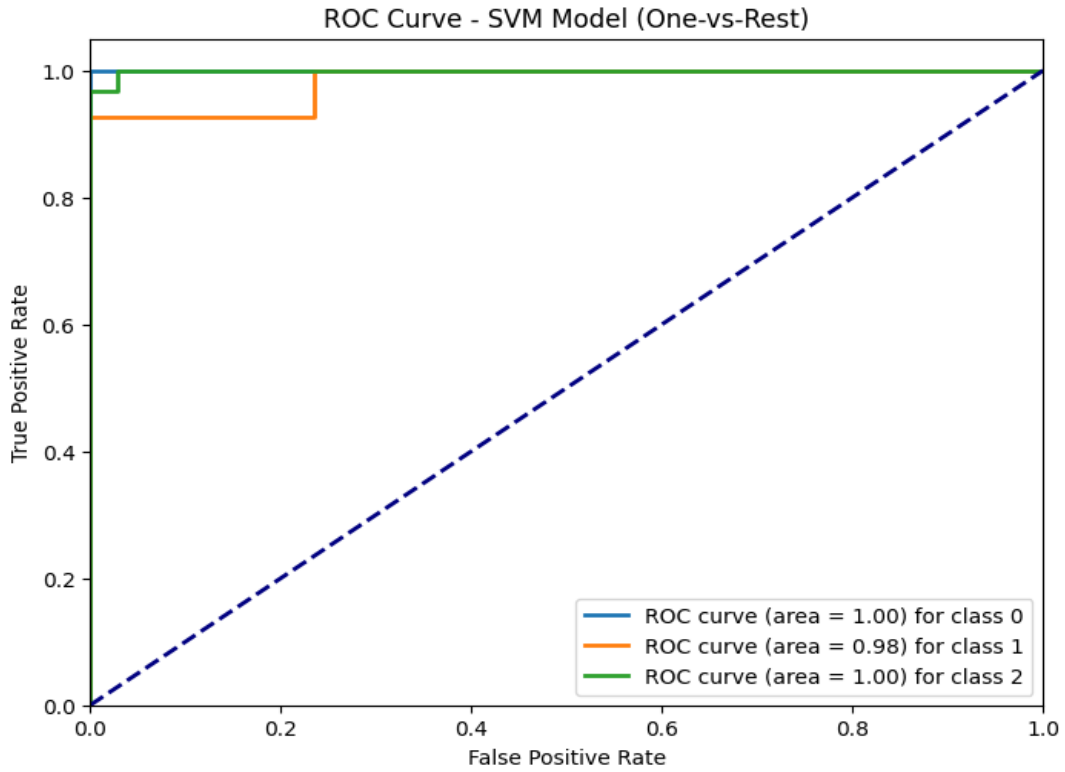


Figure 4.21: AUC-ROC Curve Analysis of SVM

SVM achieved the score, 97.50%. the precision 99%, recall 100% and F-1 score had 99%. which was a lowest score by other models.

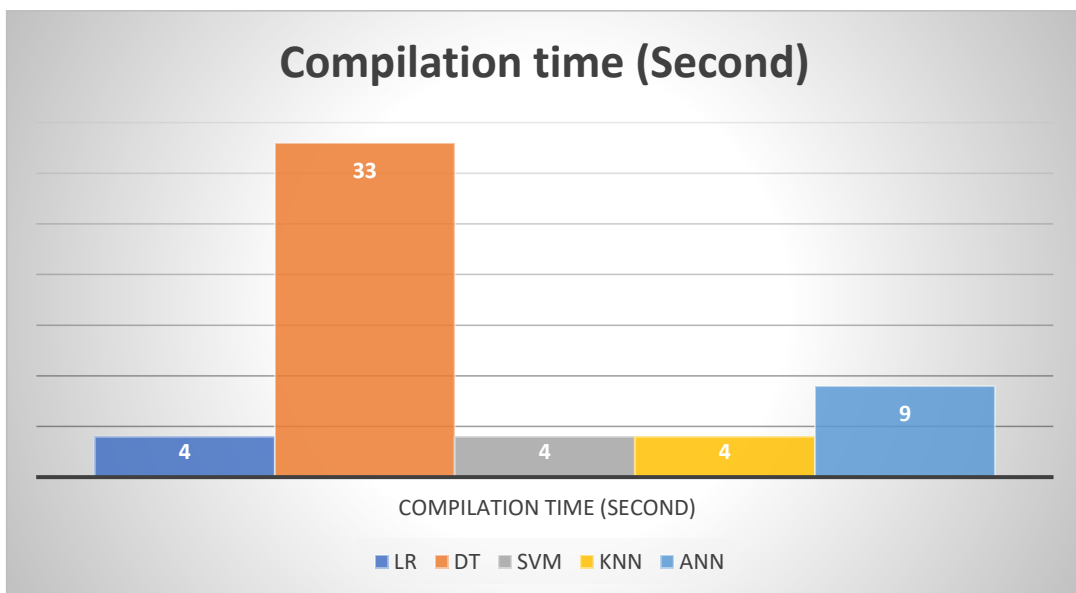


Figure 4.22: Compilation Time (Second)



I have calculated the compilation time for the suggested approach that is shown in Figure 4.19. I have calculated specific compilation times for certain algorithms. The Decision Tree method had the longest compilation time of all the algorithms, taking 33 seconds, as the picture illustrates. The Artificial Neural Network (ANN) method then required 9 seconds, which is the second-longest algorithmic compilation time.

### 4.3 Discussion

In this level, I had defined my suggested model's court system. I took into account the accuracy, precision, recall, and F-1 score.

#### 4.3.1 Accuracy

The percentage of accurate predictions generated from testing data is expressed as a statistic called accuracy. Accuracy depends on real measurements, in contrast to availability measures. It handles purposeful mistakes explicitly and is based on a single component. One of the easiest ways to evaluate any model is to look for precision. To guarantee the precision of my models, further work and improvement are needed.

$$Accuracy = \frac{TruePositive + TrueNegative}{TruePositive + FalsePositive + TrueNegative + FalseNegative}$$

#### 4.3.2 Precision

The proportion of optimistically projected observations that came to pass is measured by precision. It calculates the actual percentage of all instances that were correctly identified as true. For any kind of model, a high recall rate might be misleading, however.

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

### 4.3.3 Recall

It speaks about the proportion of data produced by the model that ought to be positive. Although great precision is ideal, there are times when it might be deceptive. The ratio of all positive labels to the predicted positives is usually established using recall.

$$\text{Recall} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalseNegative}}$$

### 4.3.4 F-1 Score

Recall and accuracy measurements are mentioned as being crucial for assessing performance. These proportions are important markers. A lower harmonic mean indicates that the model may not function as well.

$$F-1 \text{ Score} = 2 * \frac{\text{Recall Precision}}{\text{Recall} + \text{Precision}}$$

## **CHAPTER 5**

### **IMPACT ON SOCIETY, ENVIRONMENT, ETHICAL ASPECTS, AND SUSTAINABILITY**

#### **5.1 Impact on Society**

There are many social and economic benefits associated with my suggested strategy. My model carefully examines and pinpoints the essential elements or traits of a lung cancer patient by using real-world field data. The project has the potential to educate people about lung cancer prevalence and prevention strategies, which is one social benefit. Furthermore, via accurate diagnosis and regular checkups, early therapy can be prescribed, increasing the likelihood of individuals being aware of illnesses and determining their susceptibility. My method streamlines the process, requiring fewer compilations and offering quicker results, resulting in precise and unambiguous sickness prediction. By employing cutting-edge diagnostic techniques, I analyzed the data within my model to pinpoint the underlying causes of lung disease. I am optimistic that my recommended strategy will be adopted and recognized at the social level.

#### **5.2 Impact on Environment**

My suggested paradigm is especially helpful in remote places since it has fewer diagnostic steps. I can reduce the amount of effort and complexity by using the device model. We have a straightforward process that will benefit not just the environment but also ourselves. Patients do not have to go to major cities to get a lung cancer diagnosis. The patient's diagnostic report and the predictive model, which can foresee possible outcomes, might be combined. Patients won't have to worry about the cheap cost of identifying heart disease or the high cost of receiving local therapy because of its low cost. It is less complex, so people with different skill levels may use it. Our proposed method may be used to diagnose lung cancer in a patient. With my proposed paradigm, the social and economic environment will be enhanced. I am certain that my proposed approach, if accepted for use, would represent a major breakthrough in the area of medical scientific technology.

### **5.3 Ethical Aspects**

I have to include certain moral prohibitions that forbid publishing private medical information, hilarious stuff, or personal information before I can launch the system. Future research on lung cancer as well as practical lung cancer diagnosis and treatment may benefit from our suggested method. I learned that issue affects the whole globe, not just a little bit area of it. The suggested model may be used by any victim or informed individual to predict how quickly their lung cancer will progress. Processing sensitive health data from a thousand individuals raises privacy concerns. Protecting individuals' privacy requires making sure that the right data anonymization is done and that data protection guidelines are followed.

### **5.4 Sustainability Plan**

I am certain that my proposed paradigm will be accepted by the technology utilized worldwide for lung cancer research and diagnosis. I have no doubt that our recommended approach, which allows victims to evaluate their risk of acquiring lung cancer, will benefit both male and female victims. With the right tools and space, I may be motivated and ready to support rural areas. I have proposed will be beneficial and long-lasting.

## **CHAPTER 6**

### **SUMMARY, CONCLUSION, RECOMMENDATION, AND IMPLICATION FOR FUTURE RESEARCH**

#### **6.1 Summary of the Study**

I analyzed the influenced rate of our folks in our intriguing piece using algorithms. I can make accurate predictions about the future using my model. The diagnostic technology could help the prediction method. It might be useful for people to know whether they would have an impact or not. It's easy for individuals to mistakenly think that they ought to know about lung cancer. The various stages of lung cancer will be easy for people to recognize if they use my method. In light of my proposed paradigm, diagnosing authority could also be useful. Various well-known algorithms that are easy to build, very accurate, and need minimal training were used by me.

#### **6.2 Conclusion**

We inhabit a modern society characterized by technological advancement, yet simplicity remains a key aspect. The accessibility of new technology allows everyone on Earth to utilize it effectively. Leveraging these technological improvements, I aim to provide fast and easy solutions, particularly in predicting human heart disease. My objective is to simplify the process and offer assistance to our communities through innovative models. However, feasibility must be ensured before proceeding, and I am committed to incorporating additional features and addressing more prevalent issues in the future. These are the expectations I set forth.

#### **6.3 Implication for Further Study**

As mortal beings, we are susceptible to numerous illnesses on a daily basis. While lung cancer impacts many of us, some individuals are fortunate enough to recover from it. In developed nations like ours, the technology used for diagnosis and treatment is notably advanced and precise. Thanks to these technological advancements, the detection of lung cancer has become increasingly easier and faster. I have endeavored to provide our clients with a unique offering, and I am optimistic that more individuals will empathize with our

perspective. To enhance performance, I have worked on several algorithms and plan to incorporate more in the future.

#### **6.4 Limitations**

Humans are mortal beings. I come into contact with many different types of illnesses on a regular basis. Of us, most have cancer, yet some have needs related to recovery. Modern therapeutic and diagnostic technologies are evolving to become more precise and dynamic due to the rapidly changing world we live in. I'll commit more effort to this project if the public accepts our suggested approaches. The evaluation may vary from other research approaches since our model is operated on a little amount of data.

## REFERENCE

1. Chaturvedi, P., Jhamb, A., Vanani, M., & Nemade, V. (2021b). Prediction and classification of lung cancer using machine learning techniques. *IOP Conference Series*, 1099(1), 012059. <https://doi.org/10.1088/1757-899x/1099/1/012059>
2. Saleem, R., Shah, J. H., Sharif, M., & Ansari, G. J. (2021). Mango leaf disease identification using fully resolution convolutional network. *Computers, Materials & Continua*, 69(3), 3581–3601. <https://doi.org/10.32604/cmc.2021.017700>
3. Kadir, T., & Gleeson, F. (2018b). Lung cancer prediction using machine learning and advanced imaging techniques. *Translational Lung Cancer Research*, 7(3), 304–312. <https://doi.org/10.21037/tlcr.2018.05.15>
4. Adetiba, E., & Olugbara, O. O. (2015). Lung Cancer Prediction Using Neural Network Ensemble with Histogram of Oriented Gradient Genomic Features. *The Scientific World Journal*, 2015, 1–17. <https://doi.org/10.1155/2015/786013>
5. Podolsky, M. D., Барчук, A., Kuznetsov, V. I., Gusarova, N., Gaidukov, V. S., & Tarakanov, S. A. (2016). Evaluation of machine learning algorithm utilization for lung cancer classification based on gene expression levels. *Asian Pacific Journal of Cancer Prevention*, 17(2), 835–838. <https://doi.org/10.7314/apjcp.2016.17.2.835>
6. Patra, R. (2020). Prediction of lung cancer using machine learning classifier. In *Communications in computer and information science* (pp. 132–142). [https://doi.org/10.1007/978-981-15-6648-6\\_11](https://doi.org/10.1007/978-981-15-6648-6_11)
7. Morgado, J., Pereira, T., Silva, F., Freitas, C., Negrão, E., De Lima, B. F., Da Silva, M. C., Madureira, A. J., Ramos, I., Hespanhol, V., Costa, J. L., Cunha, A., & Oliveira, H. P. (2021). Machine learning and feature selection methods for EGFR mutation status prediction in lung cancer. *Applied Sciences*, 11(7), 3273. <https://doi.org/10.3390/app11073273>

7. Yakar, M., Etiz, D., Metintaş, M., Ak, G., & Çelik, Ö. (2021). Prediction of radiation pneumonitis with machine learning in stage III lung cancer: a pilot study. *Technology in Cancer Research & Treatment*, 20, 153303382110163. <https://doi.org/10.1177/15330338211016373>
8. Gould, M. K., Huang, B., Tammemägi, M. C., Kinar, Y., & Shiff, R. (2021). Machine learning for early lung cancer identification using routine clinical and laboratory data. *American Journal of Respiratory and Critical Care Medicine*, 204(4), 445–453. <https://doi.org/10.1164/rccm.202007-2791oc>
9. Δρίτσαας, Η., & Trigka, Μ. (2022). Lung Cancer Risk Prediction with Machine Learning Models. *Big Data and Cognitive Computing*, 6(4), 139. <https://doi.org/10.3390/bdcc6040139>
10. Lu, J., Tan, L., & Jiang, H. (2021). Review on Convolutional Neural Network (CNN) applied to plant leaf disease classification. *Agriculture*, 11(8), 707. <https://doi.org/10.3390/agriculture11080707>
11. International, B. R. (2023). Retracted: Lung Cancer Prediction from Text Datasets Using Machine Learning. *BioMed Research International*, 2023, 1. <https://doi.org/10.1155/2023/9790635>
12. Wu, Z., Wang, F., Cao, W., Qin, C., Dong, X., Yang, Z., Zheng, Y., Luo, Z., Zhao, L., Yu, Y., Xu, Y., Jiang, L., Tang, W., Shen, S., Wu, N., Tan, F., Li, N., & He, J. (2022). Lung cancer risk prediction models based on pulmonary nodules: A systematic review. *Thoracic Cancer*, 13(5), 664–677. <https://doi.org/10.1111/1759-7714.14333>
13. Manju, B. R., Athira, V., & Rajendran, A. (2021). Efficient multi-level lung cancer prediction model using support vector machine classifier. *IOP Conference Series*, 1012(1), 012034. <https://doi.org/10.1088/1757-899x/1012/1/012034>
14. Nam, Y., & Shin, W. (2019). A study on comparison of lung cancer prediction using ensemble machine learning. *Korean Journal of Artificial Intelligence (Online)*, 7(2), 19–24. <https://doi.org/10.24225/kjai.2019.7.2.19>



15. Azzawi, H., Hou, J., Xiang, Y., & Alanni, R. (2016). Lung cancer prediction from microarray data by gene expression programming. *Iet Systems Biology*, 10(5), 168–178. <https://doi.org/10.1049/iet-syb.2015.0082>
16. Markaki, M., Tsamardinos, I., Langhammer, A., Lagani, V., Hveem, K., & Røe, O. D. (2018). A Validated clinical risk prediction model for lung cancer in smokers of all ages and exposure types: a HUNT study. *EBioMedicine*, 31, 36–46. <https://doi.org/10.1016/j.ebiom.2018.03.027>
17. Ellen, J. G., Jacob, E., Nikolaou, N., & Markuzon, N. (2023). Autoencoder-based multimodal prediction of non-small cell lung cancer survival. *Scientific Reports*, 13(1). <https://doi.org/10.1038/s41598-023-42365-x>
18. Jabbari, F., Villaruz, L. C., Davis, M., & Cooper, G. F. (2020). Lung cancer survival Prediction using Instance-Specific Bayesian networks. In *Lecture Notes in Computer Science* (pp. 149–159). [https://doi.org/10.1007/978-3-030-59137-3\\_14](https://doi.org/10.1007/978-3-030-59137-3_14)
19. Xie, N., Hu, L., & Li, T. (2015). Lung cancer risk prediction method based on feature selection and artificial neural network. *Asian Pacific Journal of Cancer Prevention*, 15(23), 10539–10542. <https://doi.org/10.7314/apjcp.2014.15.23.10539>
20. Wu, Y., Liu, J., Han, C., Liu, X., Chong, Y., Wang, Z., Gong, L., Zhang, J., Gao, X., Guo, C., Liang, N., & Li, S. (2020). Preoperative prediction of lymph node metastasis in patients with Early-T-Stage non-small cell lung cancer by machine learning algorithms. *Frontiers in Oncology*, 10. <https://doi.org/10.3389/fonc.2020.00743>

## Lung Cancer Stage Prediction Using Machine Learning

### ORIGINALITY REPORT

<b>22%</b> SIMILARITY INDEX	<b>19%</b> INTERNET SOURCES	<b>2%</b> PUBLICATIONS	<b>8%</b> STUDENT PAPERS
--------------------------------	--------------------------------	---------------------------	-----------------------------

### PRIMARY SOURCES

<b>1</b>	<b>dspace.daffodilvarsity.edu.bd:8080</b> Internet Source	<b>12%</b>
<b>2</b>	<b>Submitted to University of Greenwich</b> Student Paper	<b>2%</b>
<b>3</b>	<b>Submitted to Daffodil International University</b> Student Paper	<b>2%</b>
<b>4</b>	<b>ieee-hpec.org</b> Internet Source	<b>&lt;1%</b>
<b>5</b>	<b>osuva.uwasa.fi</b> Internet Source	<b>&lt;1%</b>
<b>6</b>	<b>Submitted to University of New England</b> Student Paper	<b>&lt;1%</b>
<b>7</b>	<b>www.analyticsvidhya.com</b> Internet Source	<b>&lt;1%</b>
<b>8</b>	<b>Submitted to University of Leeds</b> Student Paper	<b>&lt;1%</b>
<b>9</b>	<b>Submitted to Edge Hill University</b> Student Paper	<b>&lt;1%</b>