munna

by Shariar Kabir

Submission date: 17-Jan-2022 12:34PM (UTC-0500)
Submission ID: 1742922818
File name: 160133.docx (921.23K)
Word count: 11016
Character count: 58959

LUNG CANCER PREDICTION USING MACHINE LEARNING ALGORITHMS

Department of Computer Science and Engineering

Pabna University of Science and Technology, Pabna-6600

Course Title: Thesis
Course Code: CSE 4100 and CSE 4200

A thesis has been submitted to the Department of Computer Science and Engineering for the fulfillment of the requirement of BSc Engineering Degree in Computer Science and Engineering

Submitted by:	**Supervised by:**
Md. Mohimenol Islam Id :160133 Registration Number: 101711, Session: 2015-16 Pabna University of Science and Technology.	Md. Shafiul Azam Associate Professor, Department of Computer Science and Engineering Pabna University of Science and Technology.

SUBMISSION DATE: **22th January 2022**

DECLARATION

In accordance with rules and regulations of Pabna University of Science and Technology following declarations are made:

I hereby declare that this thesis has been done by me under the supervision of Md. Shafiul Azam, Assistant Professor, Department of Computer Science and Engineering, Pabna University of Science and Technology, Pabna-6600.

I also declare that neither this thesis nor any part of this thesis has been submitted elsewhere for awarding of any degree and any material reproduced in this thesis has been properly acknowledged.

Signature of the Examinee

CERTIFICATE

I am pleased to certify that Md Mohimenol Islam, Roll Number:160133, Registration Number:101711, Session: 2015-16 has performed a thesis work entitled "Lung Cancer Prediction Using Machine Learning Algorithms" under my supervision for the requirement of the completion of course entitled 'Thesis'. So far as I concern this is an original thesis that has been carried out for one year in the Department of Computer Science and Engineering, Pabna University of Science and Technology, Pabna-6600, Bangladesh. To the best of my knowledge, this paper has not been duplicated from any other paper or submitted to elsewhere prior submission to the department.

Md. Shafiul Azam
Associate Professor,
Department of Computer Science and Engineering
Pabna University of Science and Technology, Pabna-6600.
Bangladesh.

ACKNOWLEDGEMENT

First of all I would like to admit my gratefulness to the Almighty Allah for enabling me to perform this thesis successfully. I would like to express my deepest sense of gratitude to my honorable supervisor Md. Shafiul Azam, Associate Professor, Department of Computer Science and Engineering (CSE), Pabna University of Science Technology (PUST), for his scholastic supervision, valuable guidance, adequate encouragement and helpful discussion throughout the progress of this work. I am highly grateful to him for allowing me to pursuing this study under his supervision. I am deeply thankful to honorable chairman, Dr. Md. Abdur Rahim, and all the respectable teachers of the Department of Computer Science and Engineering, Pabna University of Science Technology, Pabna-6600, Bangladesh, for their encouragement to my research work. Finally, I am much grateful to my family members especially to my parents, all of my friends and well-wishers for their encouragement and supports.

January, 2022
Author

ABSTRACT

Since in respiratory system lungs are the main organs so goodness of lungs is important for every human body. These organs supply oxygen to bloodstream from air while inhale and expel carbon dioxide form bloodstream while exhale. If lungs are effected by deadly disease called cancer then there is no more hope to recover the patient life. According to the medical statistics lung cancer is the leader in mortality rate than others cancer. Early detection of lung cancer can reduce the mortality rate and also reduce the expense of treatments. For early detection of different disease nowadays many machine learning algorithms are used. In this paper we have developed six machine learning classification model using six different classification algorithms named as Support Vector Machine (SVM), Logistic Regression (LR), Decision Tree (DT), Naïve Bayes (NB), Random Forest (RF) and K-Nearest Neighbors (KNN). We have measured the performance of every classifier model using different performance metrics including precision, recall, f1-score, AUC and ROC. We have also applied different cross validation on every classifier named as K-ford cross validation, leave one out cross validation and stratified cross validation to validate the accuracy of every classifier. Result after cross validation is almost equal to the result of without cross validation. Among six classification models we have found that Decision Tree and Random forest have 100% accuracy with and without features selection. So we say that these two classifier are capable to predict lung cancer with 100% accuracy in our dataset.

Keywords: Machine Learning, Lung Cancer, KNN, SVM, DT, LR, RF, NB.

Contents

Chapter 1: Introduction 1
1.1 Introduction 1
1.2 Motivation: 2
1.3 Objective 2
1.4 Thesis Organization 3
1.6 Discussion 3
Chapter 2: Literature Review 4
2.1 Introduction 4
2.2 Related Work 4
2.3 Research Summary 5
2.4 Lung Cancer 6
2.4.1 Risk factors 6
2.4.2 Symptoms of lung cancer 7
2.4.3 Treatments of Lung Cancer: 8
2.5. Application of machine learning 9
2.6 Discussion 11
Chapter 3: System Architecture & Data Analysis 12
3.1 Architecture Design 12
3.2 Dataset 12
3.2.1 Sample of Our Dataset: 13
3.3 Features Analysis: 14
3.4 Data type of each column: 14
3.4.1 Description of Every Features: 15
3.5 Data Pre-processing 17
3.5.1 Import Dataset 18
3.5.2 Check Duplicate or Null Value 18
3.5.3 Cases in Dataset 18
3.5.4 Correlation Heatmap 19
3.5.5 Modify and Drop some columns 20
3.6 Features Selection 20
3.6.1 Score of all columns 21
3.6.2 Selected Columns 21
3.7 Dataset Splitting 22
3.7.1 Train Data 22

3.7.2 Test Data 23
Chapter 4: Machine Learning Approach & Proposed Models 24
4.1 Machine Learning 24
4.2 Machine Learning Algorithms 24
4.2.1 Supervised learning 24
4.2.2 Semi-supervised 26
4.2.3 Unsupervised learning 26
4.2.4 Reinforcement 26
4.3 Algorithm Selection and modelling 27
4.3.1 Support vector machine 27
4.3.2 Logistic regression 28
4.3.3 Decision Tree Classifier 29
4.3.4 k-nearest Neighbors 30
4.3.5 Naïve Bayes 31
4.3.6 Random Forest Classifier 33
Chapter 5: Result & Discussion 34
5.1 Introduction 34
5.2 Performance 34
5.3 Performance Metrics 34
5.3.1 Confusion Matrix 35
5.3.3 Accuracy: 36
5.3.3 Precision: 36
5.3.4 Recall/ Sensitivity: 37
5.3.5 Specificity: 37
5.3.6 F1 Score: 37
5.3.7 ROC Curve: 38
5.4 Model Performance 38
5.4.1 Support vector machine: 38
5.4.2 Logistic Regression: 41
5.4.3 Decision Tree Classifier: 43
5.4.4 k-Nearest Neighbors: 45
5.4.5 Naïve Bayes: 48
5.4.6 Random Forest Classifier: 50
5.5 Summary of Result: 53
5.6 Discussion: 55
Chapter 6: Conclusion 56

6.1 Conclusion 56
6.2 Future Scope 56
References 57

Chapter 1: Introduction

1.1 Introduction

The lungs are the main organs of the respiratory system. Every Humans have two lungs, one on each side of their chest, one on each side of their abdomen. In mammals and the majority of other vertebrates, they are located within the thoracic cavity of the chest, or near the backbone on either side of the heart. These are responsible to extract oxygen from the air while inhale and transport it to the bloodstream, as well as to expel carbon dioxide from the bloodstream while exhale. Generally, the right lung is larger than the left lung, which shares chest space with the heart. Total weight of the two lungs about 1.3 kilograms, where the right lung is heavier than the left lung.

Lung cancer is a type of cancer that develops in lung tissues, most commonly in the cells that make up the airways. It is the cause of death from cancer including both men and women, accounting for nearly a 1/4 of all cancer deaths. Cancer is a major cause of death in the world, accounting for nearly ten million deaths in 2020. From the various list of cancer here we want to mention that only for lung cancer 2.21 million dead in 2020.

Based on the presence of lung cancer cells under such a microscope, we can split lung cancer into two main classes. One is **small cell lung cancer** and another one is **non-small cell lung cancer.** Small cell lung cancer is almost primarily associated with excessive people who smoke and is less frequent than non-small cell lung cancer. Non-small cell lung cancer is a type of lung cancer which is catch-all term for a variety of different types of lung cancer. Some examples of non-small cell lung cancers are squamous cell carcinoma, adenocarcinoma and large cell carcinoma. Treatments are chosen based on the type of lung cancer.

Conventional therapies do not cure lung cancer in the majority of patients. Cancer mortality can be reduced if cases are identified and treated as soon as possible. Cancer is the most frequently to take the medication when detected early, which can result in a higher chance to survive, less mortality rates, and less expensive treatment. These lines apply to lung cancer as well. Early detection of lung cancer reduces death rates. Major improvements in the lives of lung cancer patients can be made by detecting and treating early and avoiding care delays.

For detection of lung cancer many machine learning algorithms are used in these days. Using machine learning algorithms, we can predict the condition of lung cancer.

In this our thesis book we will apply many machine learning algorithms to predict the lung cancer based on clinical data.

1.2 Motivation:

Everybody knows that lung is one of the main important organs of human body. It is responsible to supply oxygen to bloodstream while inhale and remove carbon di oxide from bloodstream while exhale. A child takes 20-30 breathe in every hour and an adult men/woman takes 12-18 breathe per hour. So, illness of lung is dangerous for particular human body. Nowadays lung cancer is very common all over the world. So, we want to set a platform or complete system where people can check the lung cancer and get accurate result with their clinical data. Lung cancer detection of early stage can reduce the mortality rate and also reduce the expense of treatment. Without early detection there is no another best way which can cure lung cancer. Cause after early detection it is possible to give appropriate treatment.

1.3 Objective

The main objective of this thesis is that we want to predict the lung cancer condition with great accuracy. Our outputs are three types. Low Medium and High. Where output low means that the patients have no lung cancer or they are not affected by lung cancer, output medium means that the patients have lung cancer but it's in an early stage, Output high means that the patients have lung cancer which is in dangerous stage.

To predict this outcome, we have used different machine learning supervised classifier algorithms including support vector machine, logistic regression, decision tree, random forest, naïve bayes, knn etc. By the mentioned classifier we build machine learning models. After building machine learning model we check the accuracy of them using different performance metrics also check cross validation to ensuring that there is no overfitting and underfitting.

1.4 Thesis Organization

In this section of our thesis book, we discussed every chapter with summarization of their content.

CHAPTER 1: (This Chapter): In this chapter we have discussed what is lung, what is lung cancer, mortality rate of lung cancer, motivation of this thesis, objective of this Thesis.

CHAPTER 2: In this chapter we have discussed the background study of our thesis, related work, machine learning, importance of machine learning etc.

CHAPTER 3: In this chapter of our thesis book, we have discussed system architecture, system design, data collection, data pre-processing, feature selection, cross validation, data splitting etc.

CHAPTER 4: In this chapter of our thesis book, we have discussed various machine learning algorithms, our selected machine learning algorithms, why we select this particular algorithm, proposed models, and many more.

CHAPTER 5: In this chapter of our thesis book, we have discussed the various performance metrics, including confusion matrix, recall, precision, f1-score, showed results of every proposed model in tabular and graphical forms.

CHAPTER 6: In this chapter of our thesis book, we have discussed future work and conclusion of our thesis book.

REFERENCE: This is link e list where we have mentioned the related papers.

1.6 Discussion

This is an introductory chapter of our thesis book. Here we have mentioned the basic terminologies which are related to machine learning based thesis.

Chapter 2: Literature Review

2.1 Introduction

We are working on **"Lung cancer prediction using machine learning algorithms".** In this chapter of our thesis book, we have discussed the literature review of our work and we also discussed the machine learning in the field of medical science.

2.2 Related Work

There are different works have been done to predict lung cancer in previous years. In the previous research many machine learning algorithms are applied to predict, and try to get many accuracies. Generally, researchers are used different types data including images data, clinical data and omics data on their works.

In this section of our book, we want to mention some related works which are done on the same domain over the years. In the reference section of our book, we have mentioned the related paper or work. Here we have mentioned the activities of the previously proposed many machines learning models on lung cancer prediction.

The authors of paper [1] developed machine learning models to predict lung cancer. They were developed their model using 5 different machine learning algorithms including Naïve Bayes, Support vector machine, Artificial neural networks Decision Tree, Random Forest etc.

The authors of paper [2] developed machine learning models to predict lung cancer. They were developed their model using 4 different machine learning algorithms including Naïve Bayes, KNN, RBF, J48. Their dataset was collected from UCI machine learning repository. It

consists of 32 instances and it has 57 features. Among their various proposed models RBF model had most accuracy which was 81%.

The authors of paper [3] developed machine learning models to predict lung cancer. They were developed their model using artificial neural networks. Their dataset was collected from data world website. It consists of 16 features. ANN model had accuracy which was 96.67%.
The authors of paper [4] developed machine learning model to predict lung cancer. Their works was based on images. Computer Tomography (CT) images are used to detect lungs cancer. Here they were used support vector machine algorithms to classify the nodule images into malignant or benign nodules and the lung nodules into malignancy levels.

The authors of paper [5] developed machine learning models to predict lung cancer. They were developed their model using 3 different machine learning algorithms including Decision Tree, k-nearest neighbors, support vector machine. Their dataset was collected from UCI machine learning repository. It is an images dataset. Among their various proposed models SVM had 0.7940347%, Decision Tree had 0.9533073% and KNN had 0.94544323 % accuracy.
The authors of paper [6] developed machine learning models to predict lung cancer. Computer Tomography (CT) images are used to detect lungs cancer. Adaptive Canny edge detection algorithm is used to detect the edges and cancer affected areas. Neural network was used for features classification, knn used to segment cancer from lung and Bayesian Regularization Neural Network (BRNN). Bayesian Regularization Neural Network (BRNN) had 99.5% accuracy.

The authors of paper [7] and the authors of paper [8] were also developed machine learning models to predict lung cancer. In both paper they had images dataset.

2.3 Research Summary

There is many research has been done yet on lung cancer using machine learning algorithms. Every paper either used numerical data or used image data to predict or detect lung cancer. Sometimes they got high accuracy sometimes got low accuracy.

2.4 Lung Cancer

To work on our titled "Lung cancer prediction using machine learning algorithms" we have needed enough knowledge about lung cancer. In this section of our book, we are going to discuss more details about the lung cancer. Lung cancer is a type of cancer that develops in lung tissues, most commonly in the cells that make up the airways. It is the cause of death from cancer including both men and women, accounting for nearly a 1/4 of all cancer deaths.

Based on the presence of lung cancer cells under such a microscope, we can split lung cancer into two main classes. One is **small cell lung cancer** and another one is **non-small cell lung cancer.** Small cell lung cancer is almost primarily associated with excessive people who smoke and is less frequent than non-small cell lung cancer. Non-small cell lung cancer is a type of lung cancer which is catch-all term for a variety of different types of lung cancer.

2.4.1 Risk factors

Lung cancer can be exacerbated by a number of factors. Some risk factors, such as smoking, can be reduced or eliminated. Other factors, including such your family history, are unmanageable.

Risk factors of lung cancer:

Smoking. The cigarettes individuals smoke per day and the number of year individuals have smoked increase their risk of getting cancer. Giving up smoking at every age can reduce the risk of having lung cancer noticeably.

Exposure to second hand smoke. Although if individuals don't smoke, being revealed to secondhand smoke raises someone risk of developing lung cancer.

Previous radiation therapy. If someone has had chest radiotherapy for yet another particular cancer, he may be at a relatively high risk for lung cancer.

Exposure to radon gas. Radioactivity is generated normally by the collapse of uranium in soil, rock, and water, and it eventually ends up in the air you breathe. Radioactivity levels that are risky can build up in any formation, such as homes.

Exposure to asbestos and other carcinogens. Place of work inhalation and other carcinogens, such as arsenic, chromium, and nickel, can increase risk of developing cancer, especially if someone keep smoking.

Family history of lung cancer. Individuals who have a parent, sibling, or child with lung cancer are at a significantly greater risk of becoming infected.**Genes.** Genes are passed from parents to children. This includes mutated genes that prevent someone cells from repairing broken DNA and others that prevent someone body from expelling cancer-causing chemical compounds from someone system.

Make your doctor aware if you or anyone in your family has had the disease. This increases the risk of getting it. A gene passed from parents to children causes estimated 8% of lung cancers.

2.4.2 Symptoms of lung cancer

Lung cancer symptoms do not always have seemed till its disease has progressed. However, some individuals face early signs and symptoms.

These include:

- Coughing.
- Hoarseness
- Blood in phlegm or sputum expelled through coughing
- Weakness
- Wheezing
- Infections that reoccur or do not resolve
- Chest pain that worsens when you cough or laugh

Advanced lung cancer symptoms include coughing, shortness of breath, chest pain, fatigue and/or unintentional weight loss. If the cancer has spread to other areas, signs and symptoms such as bone pain, headache, muscle weakness, and/or eyelid drooping may appear.

2.4.3 Treatments of Lung Cancer:

Clinical trials Getting involved in a clinical trial may be a viable treatment option for some patients. Clinical trials are conducted to determine whether new cancer treatments are safe and effective, or if they are superior to the standard treatment. Modern day standards cancer treatments are adapted from previous clinical trials. Patients who engage in a clinical trial may receive standard care or be among the first to receive a possibly new treatment.

Lung cancer screening

It refers to evaluating a healthy person who is at high risk of developing lung cancer but has no symptoms of lung cancer in able to detect lung cancer at an earlier stage when it can be treated more efficiently. When performed in a high-quality setting, low-dose chest CT-based screening has been shown to reduce the number of people dying from lung cancer while posing acceptable risks.

Chemotherapy

It refers to the use of drugs intended to kill rapidly growing cells, such as cancer cells. Chemotherapy can be injected directly into a vein or administered through a catheter, which is a thin tube that is inserted into a large vein and left there until it is no longer needed. Some chemotherapy drugs are taken orally in the form of pills.

Radiation therapy

High - energy radiation X-rays to kill cancer cells. It can be used as a stand-alone treatment or in conjunction with chemotherapy. It oftenly plays a vital role in patients by offering relieve, airway blockage, shortness of breath, or coughing.

Radiation therapy is a "focused" treatment, which means it is designed to have the greatest effect on cancer cells while causing the least amount of harm to normal cells.

Surgery

For approach is more flexible lung cancer, surgery is still recognized the "gold standard." For patients with localized disease, extracting the tumour and bordering lung tissue provides an excellent opportunity of treatment. Thoracic surgeons who specialize in the treatment of lung cancer and other chest malignancies should perform the surgery. Your surgeon will determine whether or not a tumour can be removed. Because of their proximity to, or invasion of, vital structures, not all tumours are respectable.

Surgery may not be the best option for patients who have multiple medical problems or have poor lung function. This is defined properly by our interdisciplinary approach, which involves pulmonologists, medical oncologists, and radiation oncologists.

2.5. Application of machine learning

Machine Learning is the study of how to teach computers to learn without explicitly programming them. Machine learning is one of the most trending technologies that I have ever encountered. As the name implies, it provides the computer with a feature that makes it more human-like: the capacity to learn. Machine learning is being used actively today, possibly in many more areas than one would anticipate. In chapter 4 of our thesis book, we have discussed machine learning in details. In the following section we have discussed applications of machine learning.

1. Social Media Features

Machine learning techniques and strategies are being used by social media platforms to create some attractive and excellent features. Facebook, for example, observes and archives your events, chats, likes, and comments, as well as the time you spend on precise types of posts. Machine learning learns from prior experiences and recommends friends and pages for a profile.

2. Product Recommendations

Product suggestion is one of the most well enough and broadly used machine learning applications. Product recommendation is a prominent part of nearly every e-commerce website today, and it is an intelligent application of machine learning techniques. web applications are utilising machine learning

and artificial intelligence to monitor and control someone behaviour patterns based on previous purchases, search queries, and cart history, and afterwards make product suggestions.

3. Image Recognition

One of the most meaningful and remarkable machine learning and AI techniques is image recognition, which really is a method for categorising and sensing a feature or an object in a digital image. This technique is being used for more advanced analysis such as pattern recognition, face detection, and face recognition.

4. Sentiment Analysis

One of the most significant applications of machine learning is sentiment analysis. Sentiment analysis is a real-time machine learning application that calculates the speaker's or writer's emotion or opinion. For example, if someone writes a review or an email, a sentiment analyser will immediately determine the exact consideration and tone of the text. This sentiment analysis application can be used to analysed a review-based web application, judgment applications, and other similar applications.

5. Automating Employee Access Control

Companies are currently implementing machine learning algorithms to evaluate the level of access employees would require in different areas based on their job profiles. This is one of the most exciting machine learning applications.

6. Marine Wildlife Preservation

Machine learning algorithms are being used to create behaviour models for threatened cetaceans and other marine species, which will assist researchers in organizing and coordinating their population levels.

7. Healthcare Efficiency and Medical Services

Significant healthcare sectors are currently investigating the use of machine learning algorithms to improve management. They predict patient wait times in emergency waiting rooms across various hospital departments. The models make use of vital factors that can help define the algorithm, including such staff details at different times of day, health history, and comprehensive logs of

division chats and emergency room placement. Machine learning algorithms are also used in disease detection, therapy planning, and disease prediction. This is one of the most significant applications of machine learning.

8. Predict Cancer

Another most important use case of machine learning is cancer prediction. We know that early prediction of machine learning can reduce the mortality rates, provide idea to better treatment and also reduces the treatment const.

9. In Banking Sector

Banks are now using this the most advanced machine learning technology available to assist detect theft and safeguard accounts from hacker attacks. The algorithms determine which aspects to take into account when creating a filter to keep injury at bay. Unauthentic sites will be instantly filtered out and helped prevent from transactions.

10. Language Translation

Translating is one of most commonly used machine learning implementations. Machine learning is important in interpreting from one language to another. We are amazed by how websites can seamlessly translate from one language to another while also supplying context. The technology that enables the translation tool is known as machine translation. It has enabled users to talk with people from all over the world; without it, life would be much more difficult. It has given tourists and professional colleagues the self-belief to venture into foreign lands with the guarantee that language will no longer be a barrier.

2.6 Discussion

In this chapter we have discussed our thesis literature, details about lung cancer, application of machine learning in various areas. In the next chapter of our thesis book we have discussed the system architecture and data pre-processing.

Chapter 3: System Architecture & Data Analysis

In this chapter of our book we are going to discuss from where we collect our dataset and how we analysed the dataset and all of its attributes.

3.1 Architecture Design

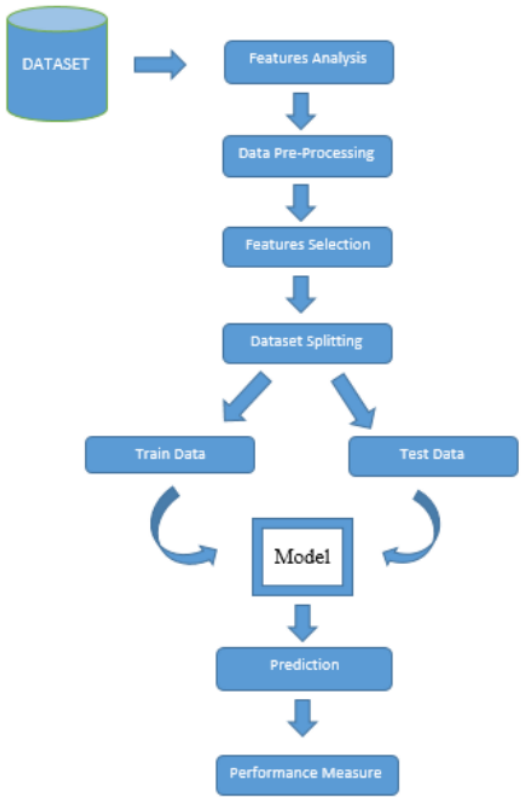


Fig 3.1 System Architecture

3.2 Dataset

The dataset which is used for our machine learning models is taken from Kaggle (https://www.kaggle.com/bilalatiq/cancerpatientsdata).

It contains 1000 records and 25 features or columns. Among 25 features 24 features are independent features and one is dependent feature. The dependent feature is the result or output column of our dataset. By 24 independent features our machine learning models will predict the lung cancer condition of a patient. All records of our data set represent a patient. Our dataset is a comma separated values (csv) file.

3.2.1 Sample of Our Dataset:

	Patient Id	Age	Gender	AirPollution	Alcoholuse	DustAllergy	OccuPationalHazards	GeneticRisk	chronicLungDisease	BalancedDiet	Obesity	Smoking	PassiveSmoker
0	P1	33	1	2	4	5	4	3	2	2	4	3	2
1	P10	17	1	3	1	5	3	4	2	2	2	2	4
2	P100	35	1	4	5	6	5	5	4	6	7	2	3
3	P1000	37	1	7	7	7	7	6	7	7	7	7	7
4	P101	46	1	6	8	7	7	7	6	7	7	8	7
5	P102	35	1	4	5	6	5	5	4	6	7	2	3
6	P103	52	2	2	4	5	4	3	2	2	4	3	2
7	P104	28	2	3	1	4	3	2	3	4	3	1	4
8	P105	35	2	4	5	6	5	6	5	5	5	6	6
9	P106	46	1	2	3	4	2	4	3	3	3	2	3

Fig 3.2: Dataset Part One

ChestPain	CoughingofBlood	Fatigue	WeightLoss	ShortnessofBreath	Wheezing	SwallowingDifficulty	ClubbingofFingerNails	FrequentCold	DryCough	Snoring	Level
2	4	3	4	2	2	3	1	2	3	4	Low
2	3	1	3	7	8	6	2	1	7	2	Medium
4	8	8	7	9	2	1	4	6	7	2	High
7	8	4	2	3	1	4	5	6	7	5	High
7	9	3	2	4	1	4	2	4	2	3	High
4	8	8	7	9	2	1	4	6	7	2	High
2	4	3	4	2	2	3	1	2	3	4	Low
3	1	3	2	2	4	2	2	3	4	3	Low
6	5	1	4	3	2	4	6	2	4	1	Medium
4	4	1	2	4	6	5	4	2	1	5	Medium

Fig 3.3: Dataset Part Two

3.3 Features Analysis:

In this section we are going to discuss about the features of our data set. In general feature or column represents an observable piece of data that can be analysed. Ex Age, Gender, Alcoholuse, ChestPain, and so on for our dataset.

Features are also known as "variables" or "attributes" in some contexts. The features include in our dataset contains the following values

1. Numerical value
2. String value

“Patient Id” and “Level” columns of our dataset have string value and others columns have numerical value. Patient id represent the unique identity for every patient. Level column represent the status of the result by storing Low, Medium and High value for every patient. Except these two feature others contain the numerical value which represent the patient health condition and based on the value of these features our models will predict the condition of lung cancer for particular patient.

3.4 Data type of each column:

```
Patient Id                object
Age                        int64
Gender                     int64
AirPollution               int64
Alcoholuse                 int64
DustAllergy                int64
OccuPationalHazards        int64
GeneticRisk                int64
chronicLungDisease         int64
BalancedDiet               int64
Obesity                    int64
Smoking                    int64
PassiveSmoker              int64
ChestPain                  int64
CoughingofBlood            int64
Fatigue                    int64
WeightLoss                 int64
ShortnessofBreath          int64
Wheezing                   int64
SwallowingDifficulty       int64
ClubbingofFingerNails      int64
FrequentCold               int64
DryCough                   int64
Snoring                    int64
Level                     object
dtype: object
```

Fig 3.4: Data Type of Each Column

3.4.1 Description of Every Features:

1. **Patient Id:**

It's represented the patient identity. It is contained character type value.

2. **Age**

It's defined the patient age. It is a numeric data.

3. **Gender**

It's defined the gender of patient. It is a Boolean data where 0 means Female and 1 means Male

4. **AirPollution**

It's contained the value of air pollution of the patient living area. It is a numeric data.

5. **Alcoholuse**

It's contained the amount of alcohole which is taken by the patient. It is a numeric data.

6. **Dust Allergy**

These features contain a value of the impact of the dust allergy of particular patient. It is a nu meric data.

7. **Occupational Hazards**

These features contain a value of the impact of the occupational hazards of particular patient. It is a numeric data.

8. **Genetic Risk**

This feature contains the value of the genetical risk of particular patient. This value comes from the patient parents. It is a numeric data.

9. **Chronic Lung Disease**

These features contain the value of the chronic lung disease of the patient. If patient have any previous lung disease, then these columns contain that value otherwise it was zero. It is a nu meric data.

10. Balanced Diet

This feature contains the value of the Balance Diet of particular patient. Balance Diet is very i mportant for human health. It is a numeric data.

11. Obesity

This feature contains the value of the obesity of particular patient. It is a numeric data.

12. Smoking

This feature contains the value of impact of smoking of particular patient. It is a numeric data .

13. Passive Smoker

This feature contains the value of impact of passive or extreme smoking of particular patient. It is a numeric data.

14. Chest Pain

This feature contains the value of impact of chest pain of particular patient. It is a numeric data.

15. Coughing of Blood

This feature contains the value of blood comes with coughing of particular patient. It is a numeric data.

16. Fatigue

The value of this field defines the tiredness of a particular patient. It is a numeric data.

17. Weight Loss

The value of this field defines the weight loss of a particular patient. It is a numeric data.

18. Shortness of Breath

The value of this field defines the value of breath shortness when a particular patient inhale. It is a numeric data.

19. Wheezing

The value of this field defines the value of wheezing depth of a particular patient. It is a numeric data.

20. Swallowing Difficulty

The value of this field defines the problem of throat of particular patient. It is a numeric value

21. Clubbing of Finger Nails

The value of this field defines the problem of throat of particular patient. It is a numeric data.

22. Frequent Cold

This field define the value of frequent cold for particular patient. It is a numeric data.

23. Dry Cough

This field define the value of dry cough for particular patient. It is a numeric data.

24. Snoring

This field define the value of snoring for particular patient. Its value is very high for some patient. It is a numeric data.

25. Level

This feature is the output column of our thesis. It contains three values, Low, Medium and High. Our proposed model predicts among these three values.

3.5 Data Pre-processing

Data Pre-processing is the initial activity for building machine learning models after collecting data. It is the process or technique of arranging raw data when used with a machine

learning model. It is indeed the most important step towards developing a machine learning model.

When we are working to develop a machine learning project, we do not always come across tidy and formatted data. And, before trying to perform any specific operation on data, it must be cleaned and formatted for the specific operation. As an outcome, we do the data pre-processing task and it is the most important and lengthy portion of working area while building machine learning model.

3.5.1 Import Dataset

To pre-process our dataset first of all we have to import dataset. For importing dataset, we use pandas. Pandas is a well-known Python-based data analysis toolkit. It comprises a wide variety of utilities, from parsing numerous file formats to converting an entire data table to a numpy matrix array. As a consequence, pandas have earned a reputation as a reliable ally in data science and machine learning.

Pandas mainly works on data in 1-D and 2-D arrays. These two arrays are handled differently in pandas.

3.5.2 Check Duplicate or Null Value

If dataset contains any null or duplicate values then this will be a concern for the accuracy of machine learning model. So, we check for null or duplicate value. Our dataset does not contain any null or duplicate value.

3.5.3 Cases in Dataset

Our Dataset contains three type of cases, Low, Medium and High. For more control over dataset, we have to know the dataset's cases ratio.

Table 3.1 Cases in Dataset

Level	Total Cases	Ratio (%)
Low	303	30.30
Medium	332	33.20
High	365	36.50

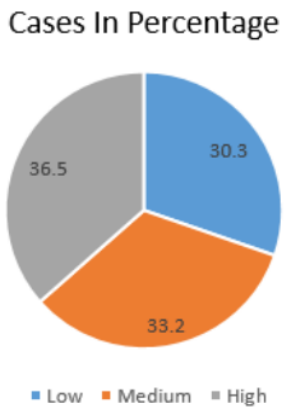


Fig 3.5 Cases Percentage

3.5.4 Correlation Heatmap

We can be defined correlation heatmap as a visual representation of the correlation matrix. It is the illustration of the relationship between different variables of the corresponding dataset. The correlation value can range from -1 to 1. By correlation heatmap we can easily visualize the relation among the features of dataset and find out the correlated columns. If there are more correlation available then the correlated column will be drop except the one form them. There are no more correlated columns in our dataset. So, we don't drop any features.

3.5.4.1 Correlation Heatmap of our dataset

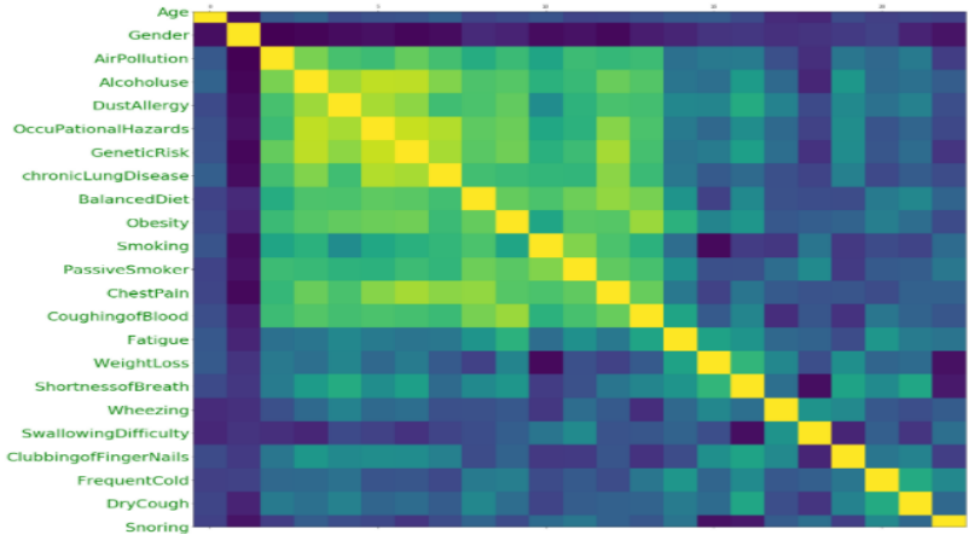


Fig 3.6: Corelation Heatmap

3.5.5 Modify and Drop some columns

In our working dataset we have a dependent column which contain the patient unique identity. This column does not have any impact on learning and accuracy of our model. So we drop this column.

We also modify the predict column which is called “level”. This column contains value in strings format. We replace the values of this columns by 0 for Low, 1 for Medium and 2 for High.

3.6 Features Selection

In practice, it is rare that all of the variables in a dataset are useful for building a machine learning model. The redundant variables decrease the model's prediction performance and may also decrease a classifier's accuracy rate. Adding more variables to train a model increases the model's overall complexity. In machine learning, feature selection technique is

used to find out the more impactful features which has more score to impact on the prediction of model. By selecting feature, total number of features are decreased. So the complexity of the model training also decrease.

Before features selection we calculate all columns score of our dataset by SelectKBest model of feature_selection of sklearn library.

3.6.1 Score of all columns

	Feature Name	Score Value
0	Age	5.752524
1	Gender	13.951461
2	AirPollution	466.785590
3	Alcoholuse	540.243082
4	DustAllergy	558.635395
5	OccuPationalHazards	413.331581
6	GeneticRisk	488.980780
7	chronicLungDisease	316.049645
8	BalancedDiet	689.937861
9	Obesity	1190.536673
10	Smoking	369.483017
11	PassiveSmoker	722.189101
12	ChestPain	404.801066
13	CoughingofBlood	1037.558618
14	Fatigue	328.928915
15	WeightLoss	97.646388
16	ShortnessofBreath	183.392939
17	Wheezing	111.397835
18	SwallowingDifficulty	44.679818
19	ClubbingofFingerNails	107.649103
20	FrequentCold	127.070943
21	DryCough	81.849012
22	Snoring	70.283627

Fig 3.7*:* Score of all columns

Form 23 dependent columns of our dataset we choose the best 18 columns which have more impact on the prediction of various machine learning model.

3.6.2 Selected Columns

Here is the list of selected columns which are more responsible to predict the lung cancer with the great accuracy.

	Feature Name	Score Value
9	Obesity	1190.536673
13	CoughingofBlood	1037.558618
11	PassiveSmoker	722.189101
8	BalancedDiet	689.937861
4	DustAllergy	558.635395
3	Alcoholuse	540.243082
6	GeneticRisk	488.980780
2	AirPollution	466.785590
5	OccuPationalHazards	413.331581
12	ChestPain	404.801066
10	Smoking	369.483017
14	Fatigue	328.928915
7	chronicLungDisease	316.049645
16	ShortnessofBreath	183.392939
20	FrequentCold	127.070943
17	Wheezing	111.397835
19	ClubbingofFingerNails	107.649103
15	WeightLoss	97.646388

Fig 3.7: Score of all selected columns

3.7 Dataset Splitting

After pre-processing the corresponding dataset, next important task is to split the dataset. In our thesis we split our dataset by the ratio of 70% of data for training and 30% of data for testing. Overfitting and underfitting are the two major factors that happen in machine learning and debase the effectiveness of the training models. To avoid the overfitting and underfitting we have needed more data for training. Another way of overcome this we use different types of cross validation. In our thesis we use following three types of cross validation to avoid the overfitting and underfitting of the model.

i. Leave one out cross-validation.
ii. K-Fold cross-validation.
iii. Stratified k-fold cross-validation

In general, we split our dataset for building machine learning model into two groups called train data and test data.

3.7.1 Train Data

Train data is the portion of our dataset which is used to train model. Model actually sees the both input and output of the train data and learn from the train data.

3.7.2 Test Data

Test Data is the portion of our dataset which is used to test the accuracy of model after training them by train data. Model don't see the output of this data. Model learn from the train data and based on the learning form train data it's predict from the test data. Outcomes from the test data we identify the accuracy of the particular model.

Chapter 4: Machine Learning Approach & Proposed Models

4.1 Machine Learning

Machine Learning is the study of how to teach computers to learn without explicitly programming them. Machine learning is one of the most trending technologies that I have ever encountered. As the name implies, it provides the computer with a feature that makes it more human-like: the capacity to learn. Machine learning is being used actively today, possibly in many more areas than one would anticipate. Machine learning is a branch of artificial intelligence that is broadly defined as a machine's ability to mimic intelligent human behaviour. Ai systems are being used to solve complex problems in a manner similar to how people deal with problems.

"Machine learning is the study of computer algorithms that allow computer programs to automatically improve through experience"

~Tom Mitchell

4.2 Machine Learning Algorithms

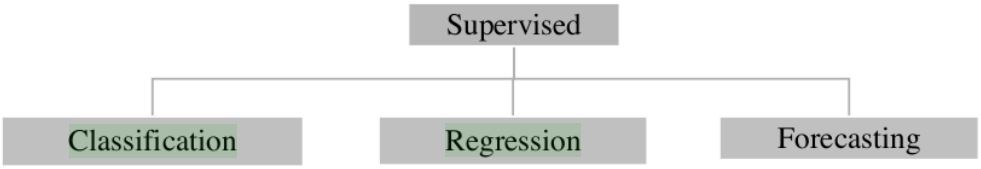
There are four types of machine learning algorithms are available.

i. Supervised
ii. Semi-supervised,
iii. Unsupervised and
iv. Reinforcement.

4.2.1 Supervised learning

The machine is trained by example in supervised learning. The operator gives the machine learning algorithm a known dataset with intended inputs and outputs, and the method must figure out how to get those inputs and outputs. While the operator is aware of the correct remedies, the algorithm recognizes data patterns, learns from observations, and makes predictions. The algorithm makes predictions, which are then corrected by the operator, and this process is repeated until the classifier accomplishes a significant amount of accuracy.

Supervised learning can be classified into three categories.



4.2.1.2 Classification

In classification supervised learning algorithms machine learning program must deduce an outcome from observations and determine which category new observations belong to. For example, predicting that patients have lung cancer or not. The program must look at existing observational data and filter the data accordingly. Classification algorithms are used when the output variable is categorical, discrete value like 1-100, 200-300 or small, medium, large.

4.2.1.2 Regression

When the input and the output features have a correlation, regression algorithms are used. It is used to forecast dependent variable such as weather, market price, and so on. The machine learning program should always estimate and comprehend the relationships among the variables in regression tasks. Regression analysis is very useful for modelling and analysis because it tends to focus on one predictor variables and a series of many other changing variables.

4.2.1.3 Forecasting

Forecasting is a method of making future predictions based on available data, and it is commonly often used discover patterns.

4.2.2 Semi-supervised

Semi-supervised learning is equivalent to supervised learning that it employs both labeled and unlabeled data. Labeled data is the data that has relevant tags so that the algorithm can acknowledge it, whereas unlabeled data does not have that information. Algorithms can gain knowledge to label and unlabeled data using this fusion. Semi-supervised machine learning is a perfect fusion of supervised and unsupervised learning methodologies. It employs a tiny amount of labeled data and a large quantity of unlabeled data, allowing it to take advantage including both unsupervised and supervised learning algorithms while attempting to avoid the difficulties associated with locating a large amount of labeled data. As a result, you can train the model to label data without ever using quite so much labeled training data.

4.2.3 Unsupervised learning

In this case, the machine learning algorithm examines data in order to identify patterns. There is no response key or human operator to provide guidance. Instead, by analysing available data, the machine determines correlations and relationships. In an unsupervised learning, the machine learning algorithm is left to perceive huge amounts of data and address that data accordingly. The algorithm attempts to organize that data in order to describe its structure. This could imply categorizing the data into clusters or organizing it in a more effective fashion. Unsupervised learning, as the name implies, is a machine learning technique in which models are not supervised using training datasets. Instead, models reveal hidden patterns and insights in the corresponding data.

4.2.4 Reinforcement

This is worried with routinized learning processes where a machine learning algorithm is offered a set of actions, parameters, and end values to implement. Following the definition of

the rules, the machine learning algorithm attempts to explore different options and possible scenarios, implementing and assessing each result to determine which is optimal. Reinforcement learning instructs the machine through trial and error. It learns from prior experience and commences to make adjustments its strategy in reaction to the situation in order to obtain the best possible outcome.

4.3 Algorithm Selection and modelling

The main goal of the entire thesis is to predict lung cancer condition with the highest accuracy. Our problem is a classification problem cause outcomes of this will be the condition of lung cancer in Low, Medium or High. In this thesis we are going to use several classification algorithms. We have chosen several algorithms typical for solving supervised learning problems throughout classification methods. In this stage we have preprocessed dataset and decided to use supervised classification algorithms.

Next step is to build our machine learning model. To build machine learning model we use sklearn library, from this library we import our selected algorithms and train them with our train data using fit function.

The various classification algorithms will be covered in the following section.

4.3.1 Support vector machine

Support vector machines are the type of supervised learning algorithm which is most used for classification, regression, and outlier detection. All those are conventional learning tasks. We can use them to detect cancerous cells using thousands of images, or we can use a well-fitted regression model to forecast future cruising routes. SVMs of different kinds can be used to solve conventional machine learning problems, including support vector regression (SVR), which is a support vector classification augmentation (SVC). The important thing to remember here is that these are simply math equations that have been tuned to give you the most precise response possible as quickly as possible.

It diverges from the other classification algorithms in that way they can select the decision boundary that maximizes the distance from the nearby data points among all classes. The maximum margin hyper plane is the decision boundary engendered by SVMs.

A simplistic linear SVM classifier connects two classes by drawing a line between them. That is, every one of the data points on one side of the line will resemble a category, while the data points on the other side of the line will be assigned to a separate category. This means that there could be an infinite number of lines to choose from. The difference of linear SVM algorithm and others including such k-nearest neighbor is that it chooses the best line to categorise the data points. It chooses the way to distinguish the data and is as much further away from the nearest data point as possible.

Types of SVMs:

There are two types of SVMs, each used for different purpose:

Simple SVM: This type of SVM is used only for linear classification and regression analysis.

Kernel SVM: This type of SVM is used for non-linear data, because it can fit a hyperplane instead of a two-dimensional space.

4.3.2 Logistic regression

This is a widely used learning algorithm under the category of supervised learning. It is used to estimate the categorical dependent feature from a set of independent features. A categorical dependent variable's output is predicted using logistic regression. As a consequence, the result must be a categorical or piecewise value. It can be Yes or No, 0 or 1. Instead of providing the accurate value it delivers the probabilistic values. Logistic Regression and Linear Regression seem to be very equivalent. Logistic regression can be used to alleviate classification types problems on the other hand linear regression algorithm can be used to remedy regression type problems.

Mechanism of logistic regression is that it does not fit a regression line, it fits a 'S' shaped logistic function that predicts two maximum values in logistic regression: 0 and 1. It is an

essential machine learning algorithm which provide probabilities and characterize new data from both continuous and discrete datasets. Logistic Regression also be used to classify observations that used a variety of data types and can quickly select the most effective features for classification. Logistic Regression estimates likelihoods using its underpinning logistic function to ascertain the correlation between the variables, the output, and the independent variables. It utilises the L2 penalty for normalization. The logistic function, also renowned as the sigmoid function, generates the resulting probabilities to binary values 0 or 1. The sigmoid function converts any real-valued number into a value between 0 and 1, with the exception of the limits themselves. Following that, a threshold classifier converts the result to a binary value. One of the core assumptions of Logistic Regression is that the input features be independent of one another.

4.3.3 Decision Tree Classifier

This is a member of the supervised learning algorithm family. Unlike other supervised learning algorithms, it can also be used to rectify regression and classification tasks. The main objective of using this algorithm is to construct a training model that predicts value of a target feature by gaining knowledge from simple decision rules from training data. In this algorithm we start from the root of the tree to predict target value for a record or row. We make comparison of the root attribute and the attributes of the record or row. It obeys the branch that pertains to that value and step to the next node analysis and the comparison.

The factor in choosing strategic splits has a substantial impact on a tree's accuracy. The decision criteria for classification and regression trees are clearly different.

To make a decision whether it should split a node into two or more sub-nodes, decision trees implement various algorithms. The emergence of sub-nodes enhances the uniform distribution of the sub-nodes that result. In other words, the integrity of the node enhances in relation to the target variable. The decision tree segregates the nodes based on all available variables and then can choose the split that yields the most uniform sub-nodes. The type of dependent variable also effects algorithm selection. Here is a list of some algorithms used throughout Decision Trees.

Types of Decision Trees:

Classification of this supervised learning algorithm is depended on the type of the target feature. There are two types of decision tree

1. Categorical Variable Decision Tree: A decision tree with such a categorical target variable is regarded to as a Categorical variable decision tree.

2. Continuous Variable Decision Tree: A decision tree with such a continuous target variable is regarded to as a Continuous variable decision tree.

4.3.4 k-nearest Neighbors

The k nearest neighbor algorithm is a Supervised Learning algorithm that is most commonly used for classification and regression. It is a powerful and flexible algorithm that can also be used to input missing values and resample datasets. As the name implies, k-nn takes into account K Nearest Data Points to predict the class or continuous value for the new data point.

Learning Types of k-nn:

i. Instance-based learning: Rather than learning weights from training data to predict output (as in model-based algorithms), we use entire training instances to predict output for previously unseen data.

ii. Lazy Learning: The model is not learned using training data before the prediction is requested on the new instance, and the learning process is postponed until the prediction is requested.

iii. Non-Parametric: There is no predetermined pattern of the mapping function in knn.

Consider the following figure. We have plotted two types sample data points from on a two-dimensional feature space. As shown, we have a total of 9 data points 5 blue circle and 4 yellow circle. Blue data points belong to ‘class A’ and yellow data points belong to ‘class B’.

And yellow star data point in a space represents the new point for which a class is to be predicted. Obviously, we say it belongs to 'class B' yellow points.

Fig 4.1 k-nearest Neighbors

This is the underlying principle of K Nearest Neighbors. In this case, nearest neighbors are those data points that are the closest in feature space to our new data point. And K is the number of such data points that we consider in our algorithm implementation. As a consequence, the distance metric and K value are two critical considerations while employing the KNN algorithm. The most broadly used distance metric is the Euclidean distance. Depending on the requirements, you may use Hamming distance, Manhattan distance, and Minkowski distance.

4.3.5 Naïve Bayes

It is indeed a classification algorithm based on Bayes' Theorem and the assumption of estimator independence. A Naive Bayes classifier, in simple terms, implies that the presence

of one feature in a class is independent of the presence of any other characteristic. The Naive Bayes model is simple to implement and is incredibly beneficial for very large amounts of

data. In additament to its easiness, Naive Bayes has been shown to outcompete even the most sophisticated classification methods.

Bayes Theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Where,

P(A|B) is Posterior probability.

P(B|A) is Likelihood probability.

P(A) is Prior Probability.

P(B) is Marginal Probability.

Types of Naive Bayes:

1. Gaussian Naïve Bayes: While characteristic values are continuous, it is hypothesised that the values associated with each class are evenly distributed according to the Gaussian distribution, also known as the Normal Distribution.

2. Multinomial Naïve Bayes: On multinomial distributed data, Multinomial Naive Bayes is preferred. It is widely used in NLP text classification. Each text classification event represents the presence of a word in a document.

3. Bernoulli Naïve Bayes: Bernoulli Naive Bayes can be used when data is spread according to multivariate Bernoulli distributions. That is, multiple features exist, but each is assumed to have a binary value.

4.3.6 Random Forest Classifier

Random forest is a versatile, user-friendly machine learning algorithm that generates outstanding results almost all of the time even without hyper-parameter tuning. For its simplicity and diversity, it is also one of the most widely used techniques. It can be used for both classification and regression tasks. We'll learn how well the random forest algorithm performs, how it deviates from other algorithms, and how to use it in this section.

Random forest is learning algorithm. It creates is an ensemble of decision trees, which have been generally trained using the "bagging" method. The stuffing method is based on the idea that trying to combine learning models enhances the overall result.

Random forest seems to have a considerable advantage such as it's used for respectively classification and regression problems, which constitute the majority of current machine learning systems. Let us just look at random forest in classification, because classification can sometimes be assumed to be the fundamental basis of machine learning. Random forest has remarkably similar hyperparameters to decision trees and bagging classifiers. Luckily, no need to merge a decision tree and a bagging classifier because of random forest classifier-class can be used instead. We can also use random forest to control regression activities while using the algorithm's regressor.

Chapter 5: Result & Discussion

5.1 Introduction

Many more tasks have been done yet, include data pre-processing, data splitting, model selection, model train etc. Now it's time to discuss about the result of our proposed models. In this chapter of our thesis book, we are going to discuss about the result of each chosen model and show the comparison of them.

5.2 Performance

Performance is the standard of a model which indicate that a particular model how much capable to predict in the corresponding domain with great accuracy. There are many ways to measure the performance of machine learning models. In the following sections we are going to discuss performances metrics to evaluate the performance of our chosen model.

5.3 Performance Metrics

Among different metrics, performance metrics are used to evaluate the performance of machine learning classification algorithms. Since our problem is a classification type problem and our selected models are classification type so we are going to use different performance metrics to evaluate the performance of our models.

5.3.1 Confusion Matrix

A Confusion matrix is a N x N matrix that is being used to examine the accuracy of the model, where N is the number of target classes. The matrix compares the actual target values to the machine learning model's predictions. This gives us with a complete view about how well our classification model is performing and the types of issues it is making.

Confusion Matrix		Predicted	
		Negative	Positive
Actual	Negative	TN	FP
	Positive	FN	TP

Fig 5.1 Confusion Matrix

1. True Positive:

The predicted value matches the actual value.

The actual value was positive and the model predicted a positive value.

2. True Negative:

The predicted value matches the actual value.

The actual value was negative and the model predicted a negative value.

3. False Positive:

The predicted value was falsely predicted.

The actual value was negative but the model predicted a positive value.

4. False Positive:

The predicted value was falsely predicted.

The actual value was positive but the model predicted a negative value.

5.3.3 Accuracy:

It is a performance metric for classification type algorithms. It may define the total number of correct predictions over the total number of predictions.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

5.3.3 Precision:

It is a performance metric for classification type algorithms. It may define how many of the correctly predicted cases were actually positive.

$$\text{Precision} = \frac{TP}{TP+FP}$$

5.3.4 Recall/ Sensitivity:

It is a performance metric for classification type algorithms. It may define how many of the actual positive cases we were able to predict correctly with our model.

$$\text{Recall/ Sensitivity} = \frac{TP}{TP+FN}$$

5.3.5 Specificity:

It is a performance metric for classification type algorithms. It is the opposite of Recall/ Sensitivity. It may define the number of negative cases we were able to predict correctly with our model.

$$\text{Specificity} = \frac{TN}{TN+FP}$$

5.3.6 F1 Score:

It is a performance metric for classification type algorithms. It is a harmonic mean of Precision and Recall.

$$\text{F1-Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

5.3.7 ROC Curve:

It is a performance metric for classification type algorithms. It is created by plotting sensitivity/recall vs 1-specificity at different threshold values. The ROC is depicted in the figure below. On the y-axis, recall/sensitivity is plotted, and on the x-axis, 1- Specificity is plotted.

5.4 Model Performance

In this section of our thesis book, we are going to discuss the performance of individual model. In the following sections we have discussed the results of every predictive model. To measure the performance or accuracy of classifier we have analysed the performance metrics including confusion matrix, precision, recall, f1-score, ROC curve etc.

5.4.1 Support vector machine:

Table 5.2 illustrate the performance metrics of support vector machine (SVM). From that table we have seen that SVM has 99.667% accuracy with all features and 100% accuracy with selected features. It has 100% precision, Recall, F1-score and AUC (area under curve). Table 5.3 illustrate the cross-validation result of SVM.

Fig 5.1 illustrate the graphical view of performance metrics; Fig 5.2 illustrate the confusion matrix and Fig 5.3 illustrate the ROC curve of this classifier.

Table 5.2 Performance Metrics:

Performance Metrics	With All Features (%)	With Selected Features (%)
Accuracy	99.667	100.00
Precession	100.00	100.00
Recall	100.00	100.00
F1-Score	100.00	100.00
AUC	100.00	100.00

Table 5.3 Cross Validation:

No	Name of Cross Validation	Accuracy (%)
01	K-Fold Cross	99.90
02	Stratified Cross	99.90
03	Leave One Out Cross	99.90

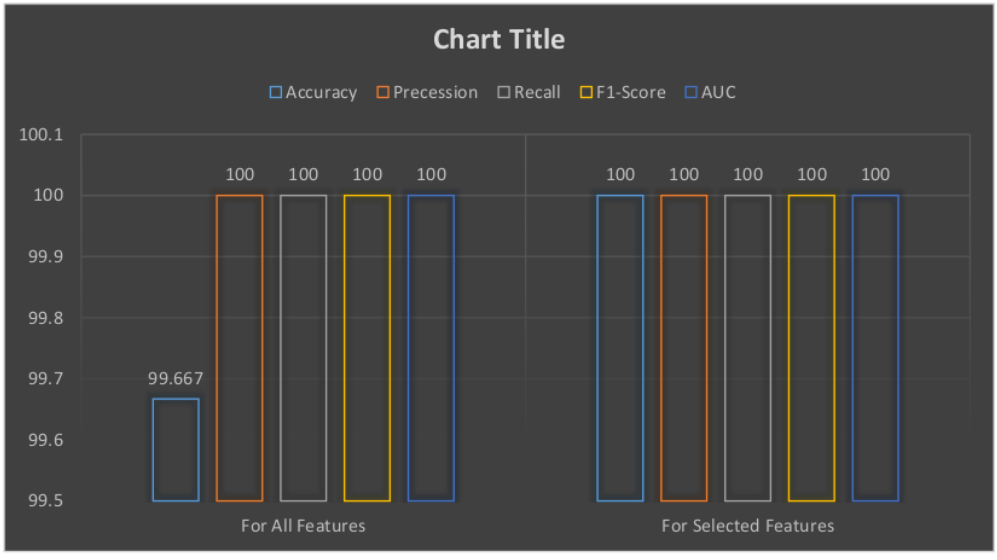


Fig 5.1: Graphical view of performance metrics

Fig 5.2: Confusion Matrix:

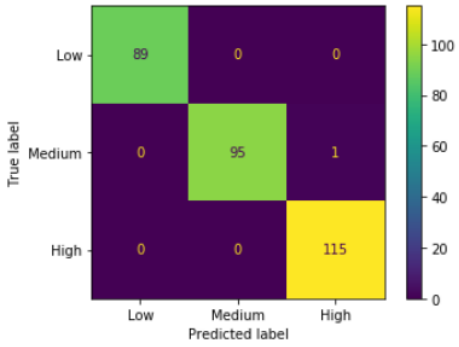


Fig 5.2(a): Confusion Matrix with All Features

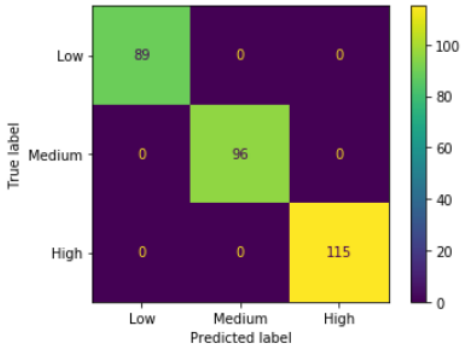


Fig 5.2(b): Confusion Matrix with selected Features.

Fig 5.3: ROC Curve:

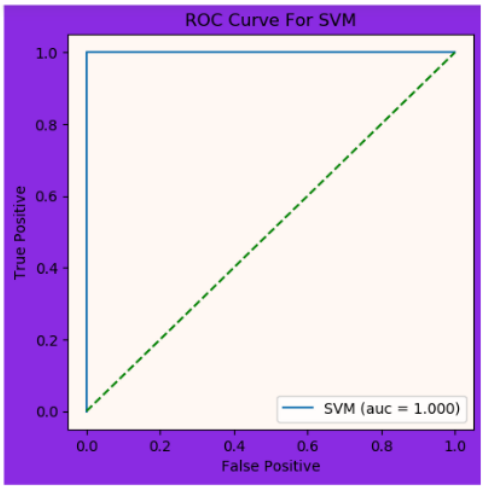


Fig 5.3(a): *Receiver operating characteristic* with All Features

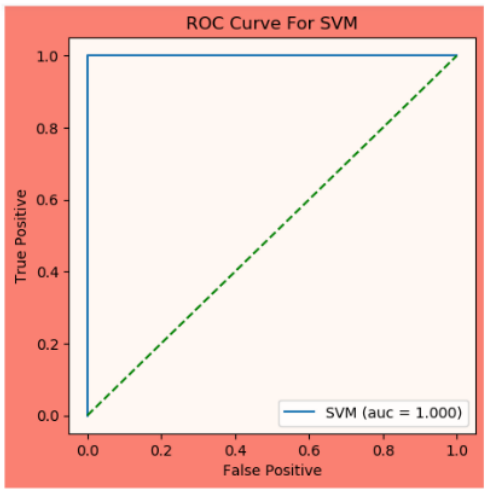


Fig 5.3(b): *Receiver operating characteristic* with selected Features.

5.4.2 Logistic Regression:

Table 5.2 illustrate the performance metrics of Logistic Regression (LR). From that table we have seen that LR has 99.667% accuracy with all features and 100% accuracy with selected features. It has 100% precision, Recall, F1-score and AUC (area under curve). Table 5.3 illustrate the cross-validation result of LR.

Fig 5.1 illustrate the graphical view of performance metrics; Fig 5.2 illustrate the confusion matrix and Fig 5.3 illustrate the ROC curve of this classifier

Table 5.4 Performance Metrics:

Performance Metrics	With All Features (%)	With Selected Features (%)
Accuracy	97.33	96.33
Precession	97.00	97.00
Recall	97.00	96.00
F1-Score	97.00	96.00
AUC	97.70	97.20

Table 5.5 Cross Validation:

No	Name of Cross Validation	Accuracy (%)
01	K-Fold Cross	98.50
02	Stratified Cross	98.50
03	Leave One Out Cross	98.40

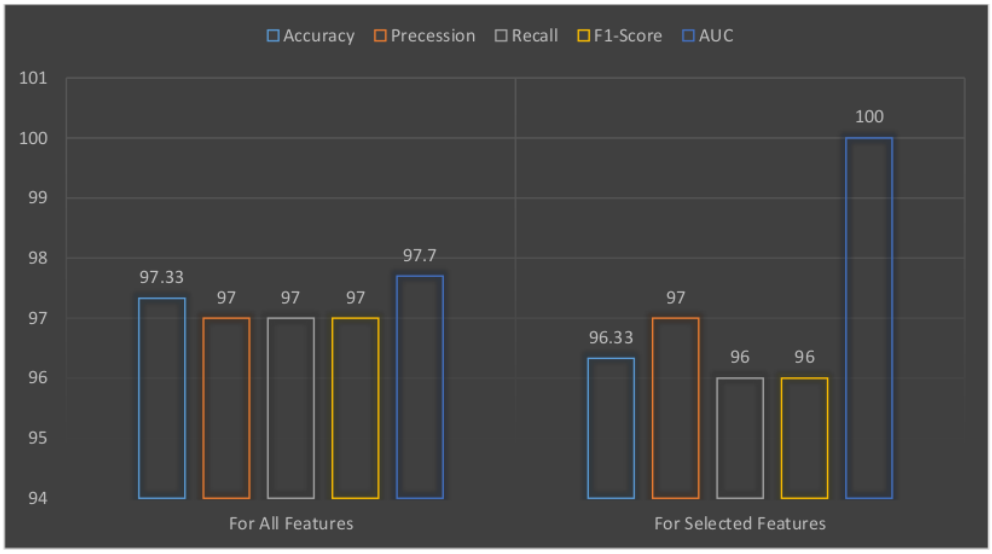


Fig 5.4: Graphical view of performance metrics

Fig 5.5: Confusion Matrix:

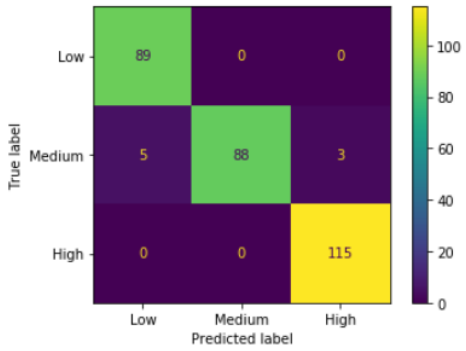


Fig 5.5(a): Confusion Matrix with All Features

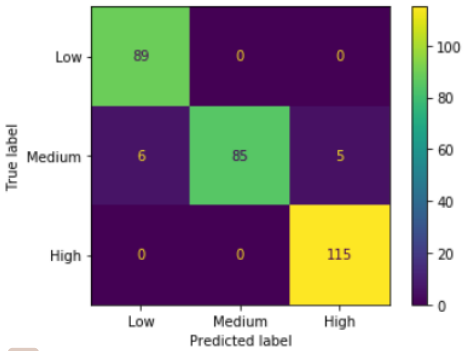


Fig 5.5(b): Confusion Matrix with selected Features.

Fig 5.6: ROC Curve:

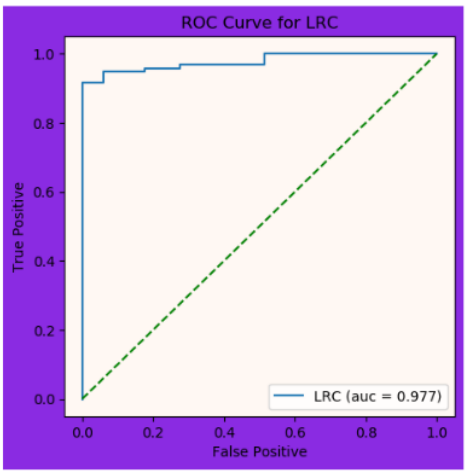


Fig 5.6(a): ***Receiver operating characteristic*** with All Features

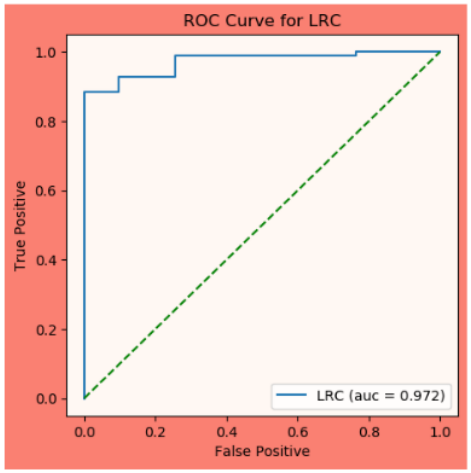


Fig 5.5(b): ***Receiver operating characteristic*** with selected Features.

5.4.3 Decision Tree Classifier:

Table 5.2 illustrate the performance metrics of Decision Tree (DT). From that table we have seen that LR has 99.667% accuracy with all features and 100% accuracy with selected features. It has 100% precision, Recall, F1-score and AUC (area under curve). Table 5.3 illustrate the cross-validation result of LR.

Fig 5.1 illustrate the graphical view of performance metrics; Fig 5.2 illustrate the confusion matrix and Fig 5.3 illustrate the ROC curve of this classifier

Table 5.6 Performance Metrics:

Performance Metrics	With All Features (%)	With Selected Features (%)
Accuracy	100.00	100.00
Precession	100.00	100.00

Recall	100.00	100.00
Specificity	100.00	100.00
F1-Score	100.00	100.00
AUC	100.00	100.00

Table 5.7 Cross Validation:

No	Name of Cross Validation	Accuracy (%)
01	K-Fold Cross	100.00
02	Stratified Cross	100.00
03	Leave One Out Cross	100.00

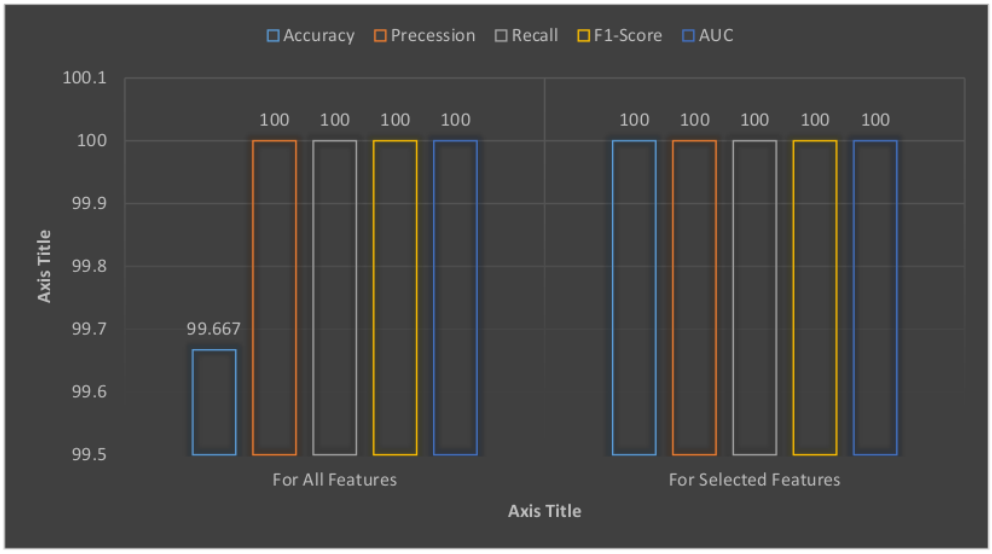


Fig 5.7: Graphical view of performance metrics

Fig 5.8: Confusion Matrix:

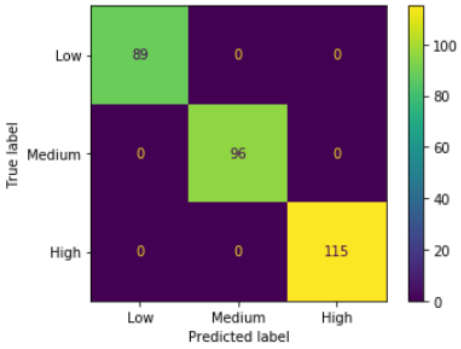


Fig 5.8(a): Confusion Matrix with All Features

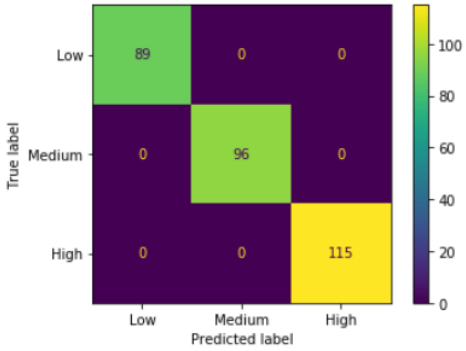


Fig 5.8(b): Confusion Matrix with selected Features.

Fig 5.9: ROC Curve:

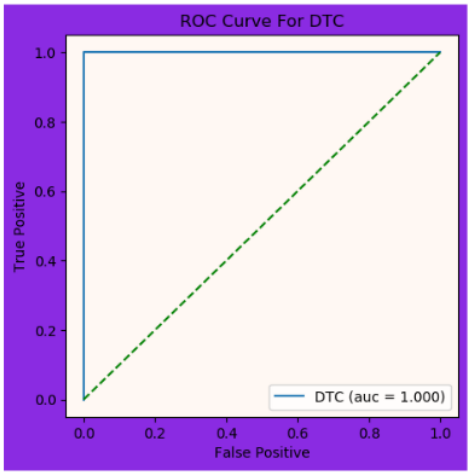


Fig 5.9(a): ***Receiver operating characteristic*** with All Features

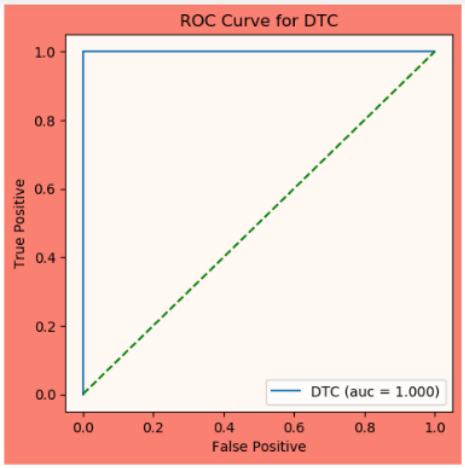


Fig 5.9(b): ***Receiver operating characteristic*** with selected Features.

5.4.4 k-Nearest Neighbors:

Table 5.2 illustrate the performance metrics of K-Nearest Neighbors (KNN). From that table we have seen that LR has 99.667% accuracy with all features and 100% accuracy with selected features. It has 100% precision, Recall, F1-score and AUC (area under curve). Table 5.3 illustrate the cross-validation result of LR.

Fig 5.1 illustrate the graphical view of performance metrics; Fig 5.2 illustrate the confusion matrix and Fig 5.3 illustrate the ROC curve of this classifier

Table 5.8 Performance Metrics:

Performance Metrics	With All Features (%)	With Selected Features (%)
Accuracy	99.667	100.00
Precession	100.00	100.00
Recall	100.00	100.00
F1-Score	100.00	100.00
AUC	99.70	100.00

Table 5.9 Cross Validation:

No	Name of Cross Validation	Accuracy (%)
01	K-Fold Cross	99.80
02	Stratified Cross	99.80
03	Leave One Out Cross	99.80

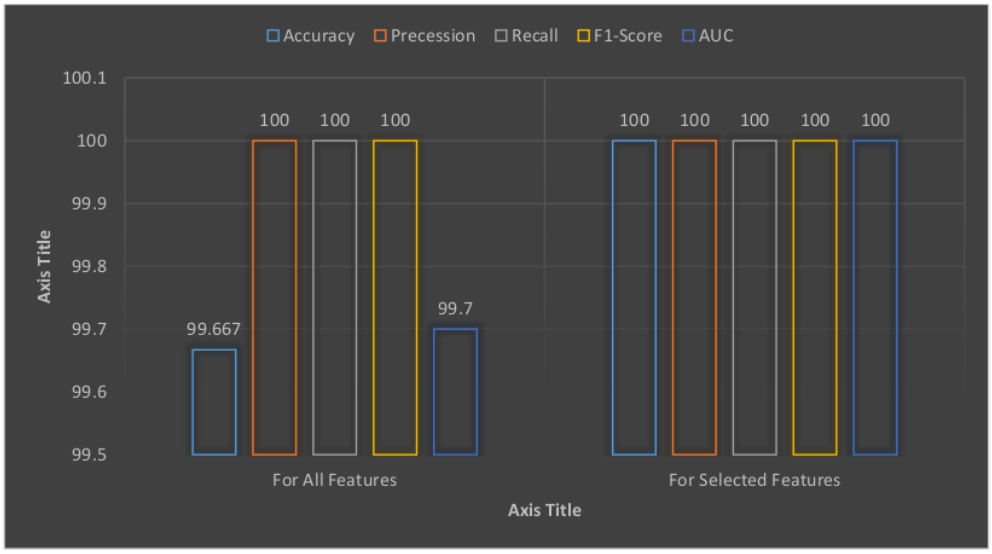


Fig 5.10: Graphical view of performance metrics

20

Fig 5.11: Confusion Matrix:

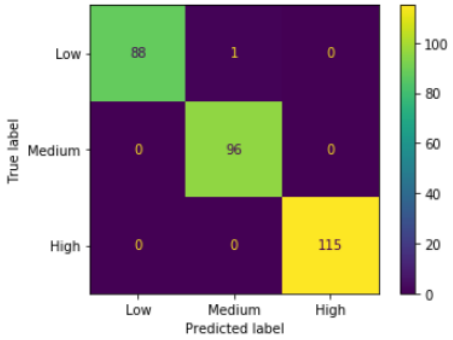


Fig 5.11(a): Confusion Matrix with All Features

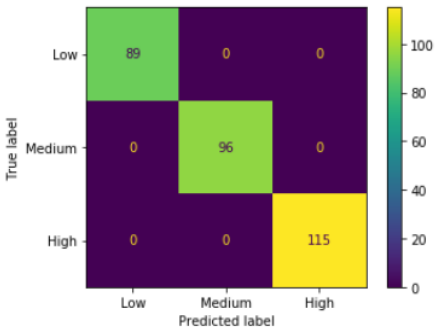


Fig 5.11(b): Confusion Matrix with selected Features.

Fig 5.12: ROC Curve:

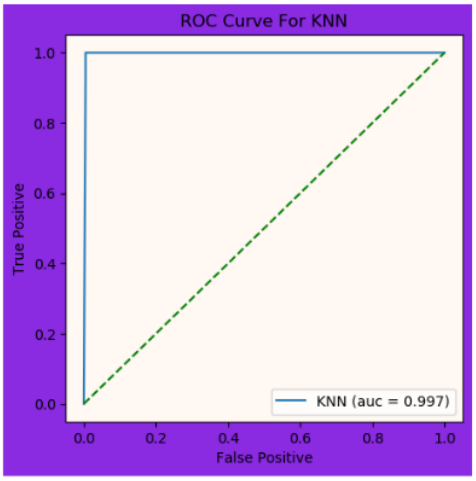


Fig 5.12(a): *Receiver operating characteristic* with All Features

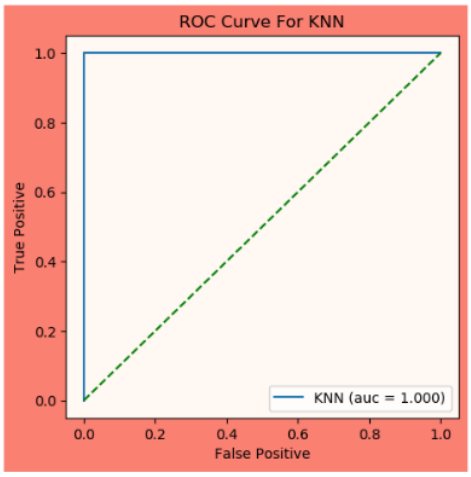


Fig 5.12(b): Receiver operating characteristic with selected Features.

5.4.5 Naïve Bayes:

Table 5.2 illustrate the performance metrics of Naïve Bayes (NB). From that table we have seen that LR has 99.667% accuracy with all features and 100% accuracy with selected features. It has 100% precision, Recall, F1-score and AUC (area under curve). Table 5.3 illustrate the cross-validation result of LR.

Fig 5.1 illustrate the graphical view of performance metrics; Fig 5.2 illustrate the confusion matrix and Fig 5.3 illustrate the ROC curve of this classifier

Table 5.10 Performance Metrics:

Performance Metrics	With All Features (%)	With Selected Features (%)
Accuracy	92.33	87.33
Precession	93.00	88.00

Recall	92.00	87.00
F1-Score	92.00	87.00
AUC	98.00	97.40

Table 5.11 Cross Validation:

No	Name of Cross Validation	Accuracy (%)
01	K-Fold Cross	89.0
02	Stratified Cross	89.1
03	Leave One Out Cross	89.0

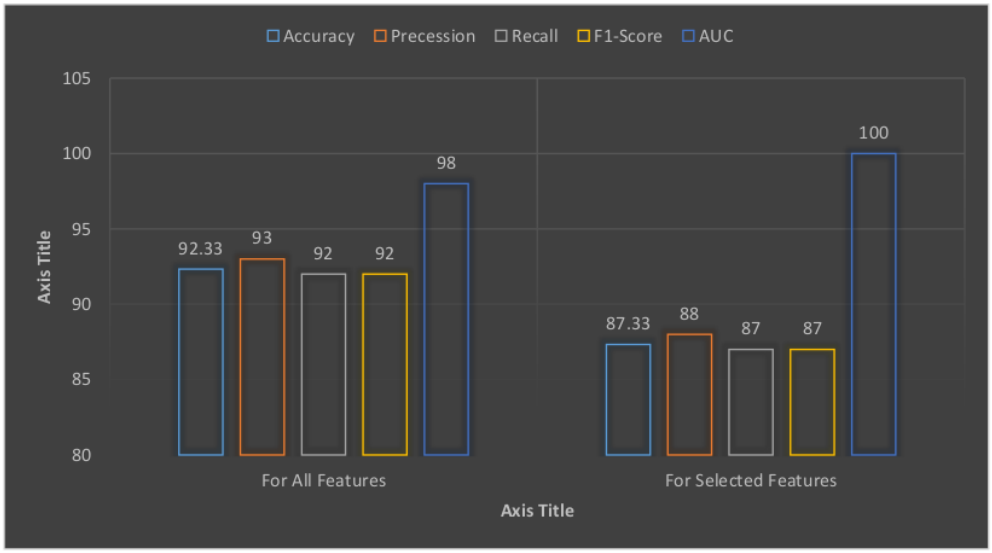


Fig 5.13: Graphical view of performance metrics

Fig 5.14: Confusion Matrix:

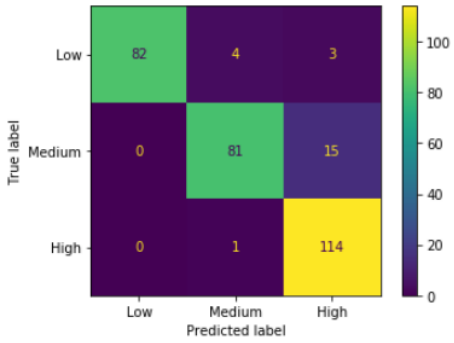


Fig 5.14(a): Confusion Matrix with All Features

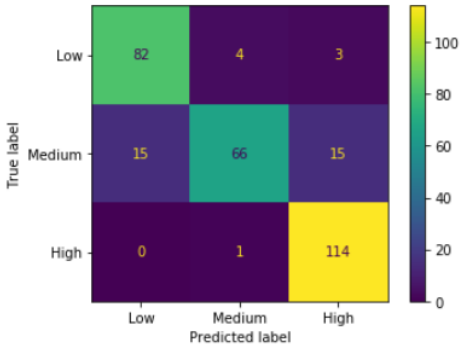


Fig 5.14(b): Confusion Matrix with selected Features.

Fig 5.15: ROC Curve:

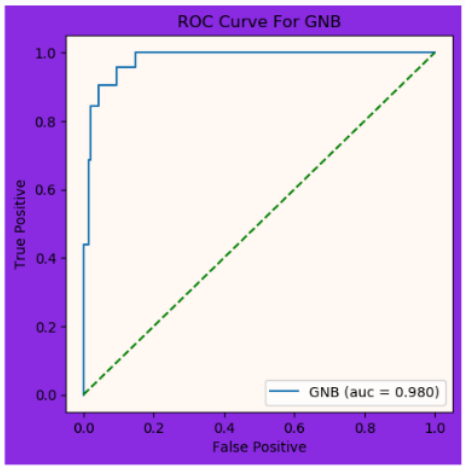


Fig 5.15(a): Receiver operating characteristic with All Features

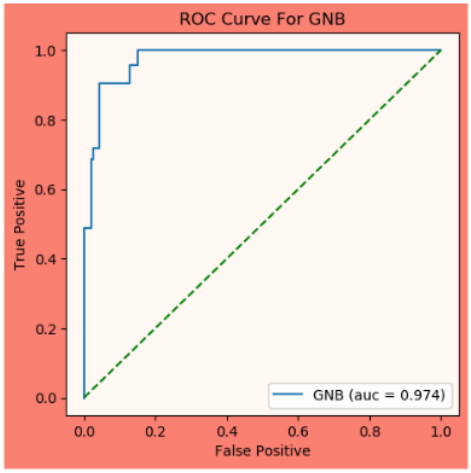


Fig 5.15(b): Receiver operating characteristic with selected Features.

5.4.6 Random Forest Classifier:

Table 5.12 illustrate the performance metrics of Random Forest Classifier (RFC). From that table we have seen that LR has 99.667% accuracy with all features and 100% accuracy with

selected features. It has 100% precision, Recall, F1-score and AUC (area under curve). Table 5.3 illustrate the cross-validation result of LR.

Fig 5.1 illustrate the graphical view of performance metrics; Fig 5.2 illustrate the confusion matrix and Fig 5.3 illustrate the ROC curve of this classifier.

Table 5.12 Performance Metrics:

Performance Metrics	With All Features (%)	With Selected Features (%)
Accuracy	100.00	100.00
Precession	100.00	100.00
Recall	100.00	100.00
F1-Score	100.00	100.00
AUC	100.00	100.00

Table 5.13 Cross Validation:

No	Name of Cross Validation	Accuracy (%)
01	K-Fold Cross	100.00
02	Stratified Cross	100.00
03	Leave One Out Cross	100.00

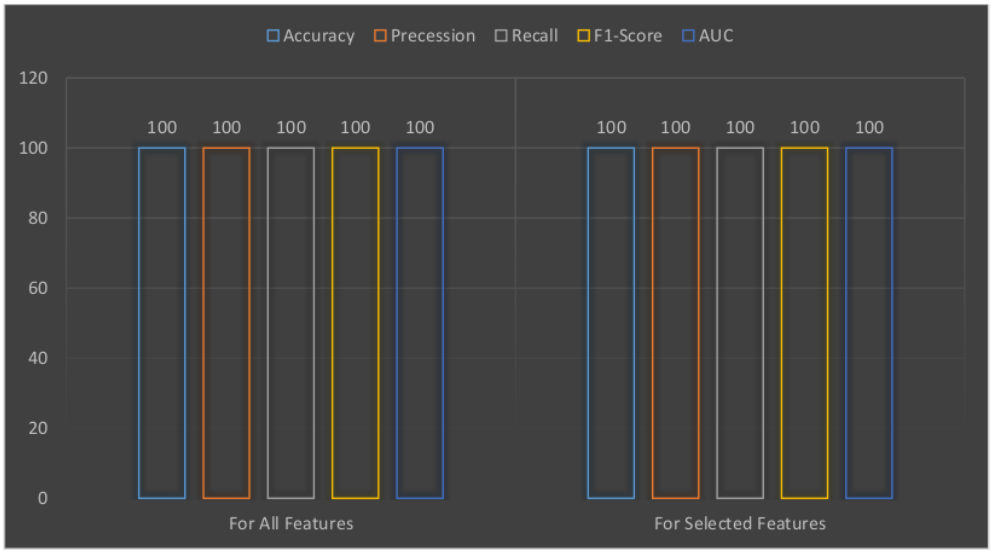


Fig 5.16: Graphical view of performance metrics

Fig 5.17: Confusion Matrix:

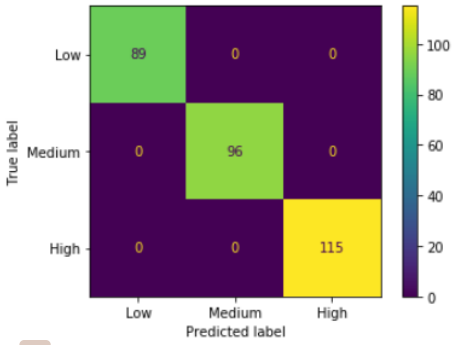


Fig 5.17(a): Confusion Matrix with All Features

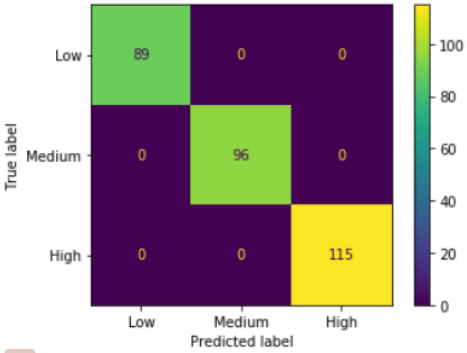


Fig 5.17(a): Confusion Matrix with selected Features.

Fig 5.18: ROC Curve:

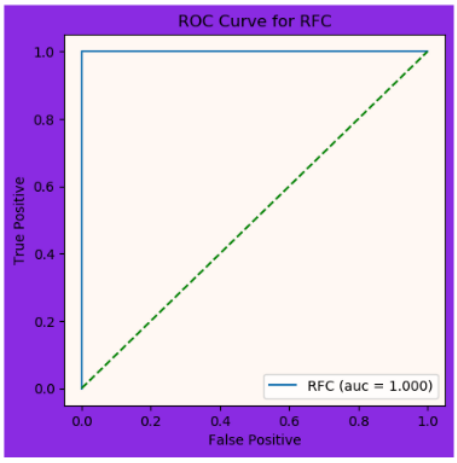


Fig 5.18(a): Receiver operating characteristic with All Features

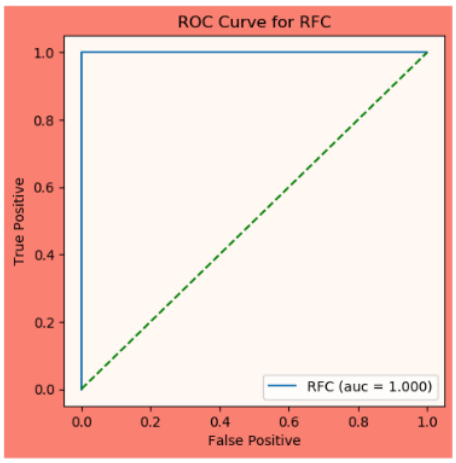


Fig 5.18(b): Receiver operating characteristic with selected Features.

5.5 Summary of Result:

Table 5.14 illustrate the performance metrics of all Classifier with all features, which we have already used to predict lung cancer. From that table we have seen that DT and RFC have 100% accuracy, and SVM, DT, KNN, RFC have 100% precision, and SVM, DT, KNN, RFC have 100% recall, and SVM, DT, KNN and RFC have 100% f1-score with all features.

Table 5.3 illustrate the performance metrics of all Classifier with selected features, which we have already used to predict lung cancer. From that table we have seen that SVM, DT, KNN and RFC have 100% accuracy, and SVM, DT, KNN, RFC have 100% precision, and SVM, DT, KNN, RFC have 100% recall, and SVM, DT, KNN and RFC have 100% f1-score with all features.

Fig 5.20 illustrate the graphical view of performance metrics and cross validation of all classifier with all features, Fig 5.21 illustrate the graphical view of performance metrics of all classifier with selected features.

Table 5.14 Result summary of all model without features selection

No	Name	Support vector machine (SVM)	Logistic regression (LR)	Decision Tree Classifier (DT)	K-Nearest Neighbors (KNN)	Naïve Bayes (NB)	Random Forest Classifier (RFC)
1	Accuracy	99.67	97.33	100.00	99.67	92.33	100.00
2	Precision	100.00	97.00	100.00	100.00	93.00	100.00
3	Recall	100.00	97.00	100.00	100.00	92.00	100.00
4	F1-Score	100.00	97.00	100.00	100.00	92.00	100.00
5	K-Fold Cross Validation	99.90	98.50	100.00	99.80	89.00	100.00
6	Stratified Cross Validation	99.90	98.50	100.00	99.80	89.10	100.00
7	Leave One Out Cross Validation	99.90	98.40	100.00	99.80	89.00	100.00
8	AUC	100.00	97.70	100.00	99.70	98.00	100.00

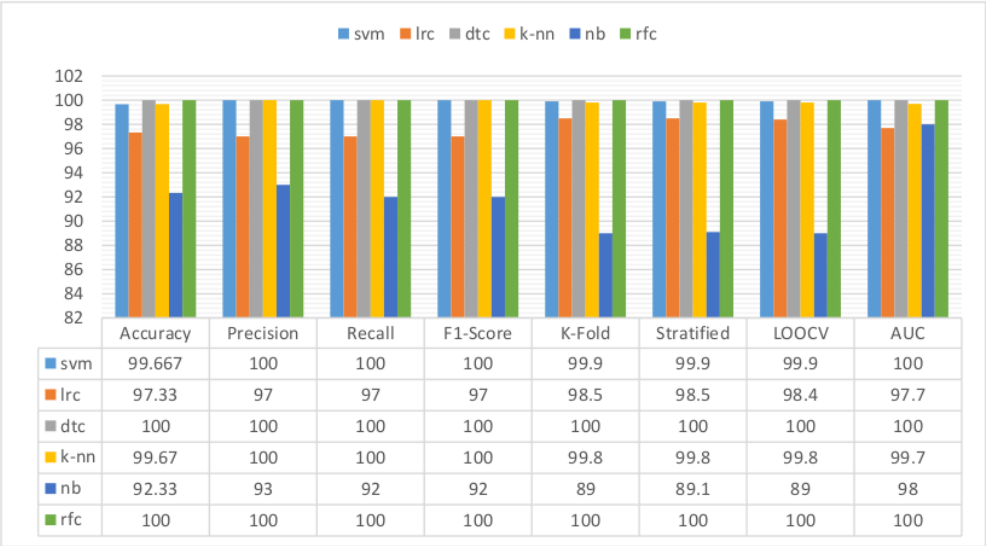


Fig: 5.19 Graphical view of result summary of all model without features selection

Table 5.15 Result summary of all model with features selection

No	Name	Support vector machine (SVM)	Logistic regression (LR)	Decision Tree Classifier (DT)	K-Nearest Neighbors (KNN)	Naïve Bayes (NB)	Random Forest Classifier (RFC)
1	Accuracy	100.00	96.33	100.00	100.00	87.33	100.00
2	Precision	100.00	97.00	100.00	100.00	88.00	100.00
3	Recall	100.00	96.00	100.00	100.00	87.00	100.00
4	F1-Score	100.00	96.00	100.00	100.00	87.00	100.00
5	AUC	100.00	97.20	100.00	100.00	97.40	100.00

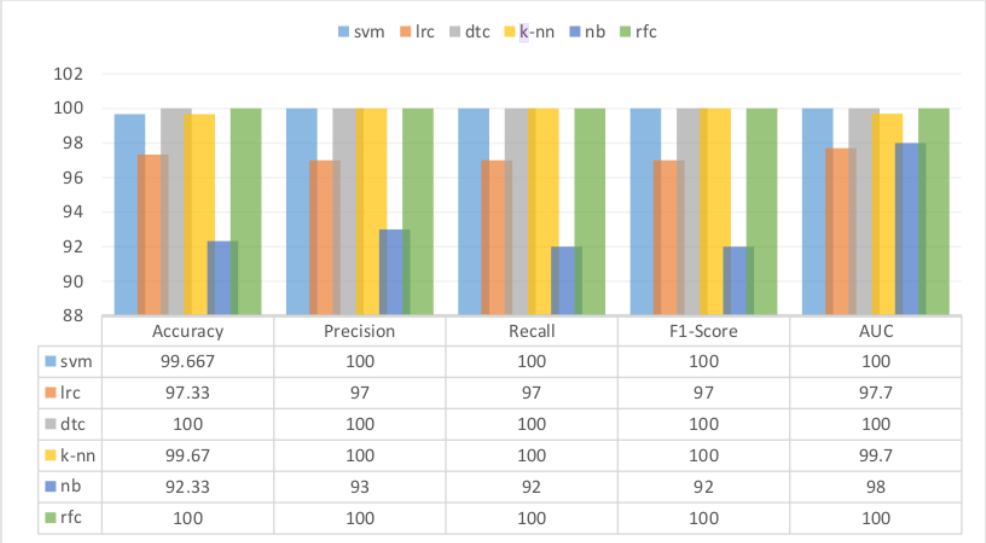


Fig: 5.20 Graphical view of result summary of all model with features selection

5.6 Discussion:

Among the above mentioned six classifier we have two classifier Decision Tree and Random Forest have 100% accuracy with all features and Decision Tree, Random Forest, Support vector machine, and K-Nearest Neighbors have 100% accuracy with selected features. So we

said that Decision Tree and Random Forest classifier can predict with 100% accuracy while features are selected or not.

Chapter 6: Conclusion

6.1 Conclusion

This is the last chapter of our thesis book. Here we have discussed the conclusion and future scope of our research.

In this research we have tried our best to develop machine learning model or we can say it a platform. By this machine learning model, a patient can test their lung cancer condition using clinical value of their medical report of human body. To build a machine learning model we have used a dataset which was full of 1000 records with 25 features or column. Its target features had three class Low, Medium and High. For effective model we have used different classifier algorithms so that we can compare our predictive result with various classifier and select the best one form them. We are very glad that we have created model that can successfully predict lung cancer with 100% accuracy. So, we hope that this model will beneficial for humans.

6.2 Future Scope

We collected our dataset from internet source. This dataset is relatively small and the records of dataset does not contain the any Bangladeshi people's data. So, our next target is to collect

Bangladeshi patient's data and perform machine learning to build predictive model. We also want to work with image data as well as clinical data.

References

[1] Dhillon, A., Kaur, A. and Singh, A., Application of Machine Learning for Prediction of Lung Cancer using Omics Data.

[2] Patra, R., 2020, March. Prediction of Lung Cancer Using Machine Learning Classifier. In *International Conference on Computing Science, Communication and Security* (pp. 132-142). Springer, Singapore.

[3] Nasser, I.M. and Abu-Naser, S.S., 2019. Lung cancer detection using artificial neural network. *International Journal of Engineering and Information Systems (IJEAIS)*, *3*(3), pp.17-23.

[4] Mohanambal, K., Nirosha, Y., Roshini, E.O., Punitha, S. and Shamini, M., 2019. Lung cancer detection using machine learning techniques. *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering*, *8*(2).

[5] Ahmed, S.R.A., Al Barazanchi, I., Mhana, A. and Abdulshaheed, H.R., 2019. Lung cancer classification using data mining and supervised learning algorithms on multi-dimensional data set. *Periodicals of Engineering and Natural Sciences (PEN)*, *7*(2), pp.438-447.

[6] Das, P., Das, B. and Dutta, H.S., 2020. *Prediction of Lungs Cancer Using Machine Learning* (No. 3076). EasyChair.

[7] Dwivedi, S.A., Borse, R.P. and Yametkar, A.M., 2014. Lung cancer detection and classification by using machine learning & multinomial Bayesian. *IOSR Journal of Electronics and Communication Engineering (IOSR-JECE)*, *9*(1), pp.69-75.

[8] Kadir, T. and Gleeson, F., 2018. Lung cancer prediction using machine learning and advanced imaging techniques. *Translational lung cancer research*, *7*(3), p.304.

munna

ORIGINALITY REPORT

32%	24%	10%	24%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Cranfield University Student Paper	2%
2	Submitted to University of Portsmouth Student Paper	2%
3	my.clevelandclinic.org Internet Source	1%
4	www.simplilearn.com Internet Source	1%
5	www.freecodecamp.org Internet Source	1%
6	archive.org Internet Source	1%
7	www.coursehero.com Internet Source	1%
8	www.analyticsvidhya.com Internet Source	1%
9	Submitted to Liverpool John Moores University Student Paper	1%

10	Submitted to International Hellenic University Student Paper	1 %
11	Submitted to University of Hertfordshire Student Paper	1 %
12	Submitted to University of Melbourne Student Paper	1 %
13	dspace.daffodilvarsity.edu.bd:8080 Internet Source	1 %
14	Submitted to Coventry University Student Paper	1 %
15	Submitted to Athlone Institute of Technology Student Paper	1 %
16	dokumen.pub Internet Source	1 %
17	ebin.pub Internet Source	1 %
18	Submitted to University of Wales Institute, Cardiff Student Paper	1 %
19	Submitted to University of Abertay Dundee Student Paper	1 %
20	Submitted to Universidad Carlos III de Madrid Student Paper	<1 %
21	Submitted to University of Bradford Student Paper	

		<1%
22	Submitted to Indian Institute of Management, Nagpur Student Paper	<1%
23	Submitted to University of Leeds Student Paper	<1%
24	www.sas.com Internet Source	<1%
25	Submitted to Goa Institute of Management Student Paper	<1%
26	lungfoundation.com.au Internet Source	<1%
27	Submitted to Institute of Graduate Studies, UiTM Student Paper	<1%
28	ijsab.com Internet Source	<1%
29	Submitted to Middle East College of Information Technology Student Paper	<1%
30	Submitted to International Business School Student Paper	<1%
31	Hossein Malekmohamadi, Nontawat Pattanjak, Roeland Bom. "Chapter 5 Human	<1%

	Activity Identification in Smart Daily Environments", Springer Science and Business Media LLC, 2020 Publication	
32	Submitted to S.P. Jain Institute of Management and Research, Mumbai Student Paper	<1 %
33	Submitted to University of Wolverhampton Student Paper	<1 %
34	www.ijeast.com Internet Source	<1 %
35	www.medicinenet.com Internet Source	<1 %
36	Submitted to MLC School Student Paper	<1 %
37	Submitted to Marquette University Student Paper	<1 %
38	Submitted to Natonal Institute of Technology Calicut Student Paper	<1 %
39	Submitted to The Scientific & Technological Research Council of Turkey (TUBITAK) Student Paper	<1 %
40	link.springer.com Internet Source	<1 %

41	Submitted to University of Energy and Natural Resources Student Paper	<1%
42	etd.lib.nsysu.edu.tw Internet Source	<1%
43	"Ambient Communications and Computer Systems", Springer Science and Business Media LLC, 2019 Publication	<1%
44	lib.buet.ac.bd:8080 Internet Source	<1%
45	Submitted to University of West Florida Student Paper	<1%
46	Wassim Bouachir, Rafik Gouiaa, Bo Li, Rita Noumeir. "Intelligent video surveillance for real-time detection of suicide attempts", Pattern Recognition Letters, 2018 Publication	<1%
47	turkjphysiotherrehabil.org Internet Source	<1%
48	Submitted to Asia Pacific University College of Technology and Innovation (UCTI) Student Paper	<1%
49	dspace.bracu.ac.bd:8080 Internet Source	<1%

50	Maria Patricia Peeris.T, P. Brundha. "Proposing an Early Diagnostic Deep Learning Approach to Detect Lung Cancer from Short-Breaths", 2019 International Conference on Smart Systems and Inventive Technology (ICSSIT), 2019 Publication	<1 %
51	byjus.com Internet Source	<1 %
52	home.simula.no Internet Source	<1 %
53	"Smart Trends in Information Technology and Computer Communications", Springer Science and Business Media LLC, 2016 Publication	<1 %
54	Submitted to University of Auckland Student Paper	<1 %
55	Submitted to Vaal University of Technology Student Paper	<1 %
56	"Computing Science, Communication and Security", Springer Science and Business Media LLC, 2020 Publication	<1 %
57	Murillo B. Rodrigues, Raul Victor M. Da Nobrega, Shara Shami A. Alves, Pedro Pedrosa Reboucas Filho et al. "Health of	<1 %

	Things Algorithms for Malignancy Level Classification of Lung Nodules", IEEE Access, 2018 Publication	
58	Submitted to Columbia University Student Paper	<1 %
59	Submitted to University of Nevada, Las Vegas Student Paper	<1 %
60	bestessays.approvedscholars.com Internet Source	<1 %
61	cdr.lib.unc.edu Internet Source	<1 %
62	Submitted to Ahsanullah University of Science and Technology Student Paper	<1 %
63	Submitted to University of Wollongong Student Paper	<1 %
64	"Proceedings of the International Conference on Big Data, IoT, and Machine Learning", Springer Science and Business Media LLC, 2022 Publication	<1 %
65	Submitted to Arab Academy for Science, Technology & Maritime Transport CAIRO Student Paper	<1 %
66	Submitted to Babes-Bolyai University	

	Student Paper	<1%
67	Submitted to Brown Mackie College Student Paper	<1%
68	Rahman, Md Mijanur, Md Farukuzzaman Khan, and Mohammad Ali Moni. "Speech recognition front-end for segmenting and clustering continuous Bangla speech", Daffodil International University Journal of Science and Technology, 2010. Publication	<1%
69	Submitted to Delhi Metropolitan Education Student Paper	<1%
70	Submitted to Higher Education Commission Pakistan Student Paper	<1%
71	Submitted to London School of Economics and Political Science Student Paper	<1%
72	doaj.org Internet Source	<1%
73	Submitted to Galgotias University, Greater Noida Student Paper	<1%
74	openaccess.hku.edu.tr Internet Source	<1%

75	Submitted to Ain Shams University Student Paper	<1%
76	Submitted to De Montfort University Student Paper	<1%
77	Submitted to Universiti Teknologi Petronas Student Paper	<1%
78	scholars.uow.edu.au Internet Source	<1%
79	text-id.123dok.com Internet Source	<1%
80	Submitted to Open Learning Group Student Paper	<1%
81	Submitted to University of Dundee Student Paper	<1%
82	bigdatawg.nist.gov Internet Source	<1%
83	dspace.nehu.ac.in Internet Source	<1%
84	www.ijser.org Internet Source	<1%
85	Submitted to Colorado Technical University Online Student Paper	<1%
86	mts.intechopen.com	

	Internet Source	<1%
87	uir.unisa.ac.za Internet Source	<1%
88	"Smart Techniques for a Smarter Planet", Springer Science and Business Media LLC, 2019 Publication	<1%
89	Submitted to Florida International University Student Paper	<1%
90	Submitted to Vilnius Gediminas Technical University Student Paper	<1%
91	digitalcommons.usu.edu Internet Source	<1%
92	docshare.tips Internet Source	<1%
93	globaljournals.org Internet Source	<1%
94	hcis-journal.springeropen.com Internet Source	<1%
95	iieta.org Internet Source	<1%
96	www.javatpoint.com Internet Source	<1%

97	"Advances in Distributed Computing and Machine Learning", Springer Science and Business Media LLC, 2022 Publication	<1 %
98	"Machine Learning and Data Mining in Pattern Recognition", Springer Nature, 2018 Publication	<1 %
99	Akshitha Shetty, Vrushika Shah. "Survey of Cervical Cancer Prediction Using Machine Learning: A Comparative Approach", 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2018 Publication	<1 %
100	Ralf Vieira de Araujo. "Machine learning for prediction of soil carbon stock changes in sugarcane crop due to straw removal", Universidade de Sao Paulo, Agencia USP de Gestao da Informacao Academica (AGUIA), 2021 Publication	<1 %
101	eprints.uthm.edu.my Internet Source	<1 %
102	jamanetwork.com Internet Source	<1 %
103	lml.bas.bg Internet Source	<1 %

104	tudr.thapar.edu:8080 Internet Source	<1%
105	www.cil.pku.edu.cn Internet Source	<1%
106	www.enggjournals.com Internet Source	<1%
107	"Computer Vision and Machine Intelligence in Medical Image Analysis", Springer Science and Business Media LLC, 2020 Publication	<1%
108	Carla Ferreira do Nascimento. "Envelhecendo na cidade: análises longitudinais do declínio da mobilidade e sobrevida de idosos com múltiplas estratégias", Universidade de Sao Paulo, Agencia USP de Gestao da Informacao Academica (AGUIA), 2021 Publication	<1%
109	www.ubuntupit.com Internet Source	<1%
110	Emilio Garcia-Fidalgo, Alberto Ortiz. "Methods for Appearance-based Loop Closure Detection", Springer Science and Business Media LLC, 2018 Publication	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		

munna

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10

PAGE 11

PAGE 12

PAGE 13

PAGE 14

PAGE 15

PAGE 16

PAGE 17

PAGE 18

PAGE 19

PAGE 20

PAGE 21

PAGE 22

PAGE 23

PAGE 24

PAGE 25

PAGE 26

PAGE 27

PAGE 28

PAGE 29

PAGE 30

PAGE 31

PAGE 32

PAGE 33

PAGE 34

PAGE 35

PAGE 36

PAGE 37

PAGE 38

PAGE 39

PAGE 40

PAGE 41

PAGE 42

PAGE 43

PAGE 44

PAGE 45

PAGE 46

PAGE 47

PAGE 48

PAGE 49

PAGE 50

PAGE 51

PAGE 52

PAGE 53

PAGE 54

PAGE 55

PAGE 56

PAGE 57

PAGE 58

PAGE 59

PAGE 60

PAGE 61

PAGE 62

PAGE 63

PAGE 64

PAGE 65

PAGE 66