LUNG CANCER DETECTION USING MACHINE LEARNING

N.Vijay Kumar[1], A.Akhila[2] , G.Tejaswini[3], G.Lilly[4]
[1] Professor, [2, 3 & 4] Student
nvk20022001@gmail.com[1] ,[2]akulaakhila13@gmail.com, [3]tejaswinigajula2003@gmail.com, [4] lillyganugapati@gmail.com
Department of Computer Science and Engineering,
Narasaraopeta Engineering College, Narasaraopet, Andhra Pradesh, India

Abstract : **Lung pain is one of the most common complaints during the early stages of cancer therapy, and the most difficult aspect is waiting for the radiologist's diagnosis. For radiologists, a high-tech computerized system is definitely very helpful. Many machine learning-based studies are available for lung cancer diagnosis. Lung cancer is strongly predicted using a multistage bracket. For data enhancement, the double classifier and threshold and marke controlled watershed used in the bracket system's segmentation system are applied. Lung cancer is an extremely difficult disease to identify. Techniques like Support Vector Machine, K Nearest Neighbour, Decision Tree, Logistic Regression, Naïve Bayes, and Random Forest are used to train the dataset, and it is shown that these algorithms show sophisticated delicacy. The best accuracy of 96.1% is obtained by K-Fold Validation with SVM and Logistic Regression, according to a comparative study of the experimental data.**

Keywords— Cancer, Machine learning Support Vector Machine, Random Forest, Logistic Regression, Naïve Bayes, K Nearest Neighbour, Decision Tree.

I. INTRODUCTION

Lung cancer is currently the most deadly cancer-related cause of death worldwide, surpassing the rates of breast, prostate, and colon cancers. Patient outcomes and death rates are greatly impacted by prompt diagnosis and treatment. But due to flaws in current screening protocols, lung cancer is frequently discovered at an advanced stage, which is associated with a worse prognosis. The use of machine learning techniques holds promise for early detection of lung cancer. These techniques can accurately estimate a person's chance of acquiring lung cancer by detecting pertinent patterns and risk variables through the analysis of large patient data.

In this paper suggest investigating machine learning on text datasets produced from electronic health records (EHRs), whereas established methods such as [1] CT-SCAN and X-RAYS concentrate on expensive and time-consuming medical imaging data. These datasets include important patient data, including demographics, medical histories, and clinical notes, which may be used to create precise risk prediction models. This study also attempts to assess earlier studies that examined genomic expression data in cancer. To identify genes linked to the development of cancer, a variety of techniques have been used, including Confusion Matrix, Accuracy, Precision, Recall, and F1 Score; additionally, algorithms such as Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Naive Bayes have been used. The foundation for creating prediction models for cancer diagnosis is these expression profiles.

II. LITERATURE SURVEY

Certainly Lung cancer detection has been a significant area of research, and there have been numerous studies focusing on this topic using various statistical and machine learning techniques. Here are some notable literature surveys for lung cancer detection:

The study provides a comprehensive global analysis of lung cancer research output and trends. It evaluates the volume and scope of lung cancer research across various countries, highlighting the disparities in research focus, funding, and output in relation to the disease's burden worldwide. The study underscores the need for increased prioritization of lung cancer research, especially in areas such as early diagnosis, screening, and palliative care, to improve clinical outcomes. The analysis also discusses the role of international collaboration and the impact of media and funding on research directions, advocating for a more equitable distribution of resources and attention to lung cancer research globally.

The research conducted by Kumar et al .offers[3] a literature survey on lung cancer prediction using machine learning techniques applied to textual datasets. They investigate the existing methodologies and advancements in leveraging machine learning for this task. The survey covers topics such as feature extraction from text, various machine learning algorithms utilized for prediction, and the integration of deep learning techniques. Additionally, the authors likely discuss challenges in the field, such as dataset size, class imbalance, and model interpretability, as well as potential avenues for future research.

The research conducted by Kumar et al. (2022) offers[3] a literature survey on lung cancer prediction using machine learning techniques applied to textual datasets. They investigate the existing methodologies and advancements in leveraging machine learning for this task. The survey covers topics such as feature extraction from text, various machine learning algorithms utilized for prediction, and the integration of deep learning techniques. Additionally, the authors likely discuss challenges in the field, such as dataset size, class imbalance, and model interpretability, as well as potential avenues for future research.

International Journal of Advanced Trends in Computer Science and Engineering, Pawar et al. (2020) present[4] a comprehensive approach to lung cancer detection by integrating image processing and machine learning techniques. The study explores the potential of combining these two domains to enhance the accuracy and efficiency of images to extract relevant features followed by classification using machine learning algorithms. The paper highlights the significance of such integrated approaches in improving early detection rates and subsequently enhancing patient outcomes.

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

This work contributes to the growing body of literature focusing on leveraging advanced technologies for more effective cancer diagnosis and treatment planning.

III.METHODOLOGY

Initially, a dataset containing patient data such as smoking status, yellow_fingers, and other relevant factors like age, gender would be collected. The dataset would be preprocessed to make sure that the data is clean, well structured, and has a balance between positive and negative cases of lung cancer. The model would be trained using the dataset and tuned using cross-validation techniques to ensure that it is not overfitting.

The methodology for lung cancer detection in machine learning involves collecting patient data like smoking status and relevant factors, preprocessing for cleanliness and balance, training the model with cross-validation, evaluating with the metrics F1-score, accuracy, precision, and recall, and comparing with other models in the literature.

Proposed model:
A. Dataset Analysis
B. Data Visualization
C. Preprocessing Techniques
D. Classification of Data
E. Creation and Evaluation

A. Dataset Analysis

The dataset [5] are sourced from Kaggle.com. and include 303 records and 16 attributes like age, gender, smoking, anxiety, yellow_fingers, shortness of breath, coughing, allergy, peer_pressure, chronic_disesse and some other attributes as shown in the below figure 1.

GENDER	AGE	SMOKING	YELLOW_F	ANXIETY	PEER_PRE	CHRONIC	FATIGUE	ALLERGY
M	69	1	2	2	1	1	2	1
M	74	2	1	1	1	2	2	2
F	59	1	1	1	2	1	2	1
M	63	2	2	2	1	1	1	1
F	63	1	2	1	1	1	1	1
F	75	1	2	1	1	2	2	2
M	52	2	1	1	1	1	2	1
F	51	2	2	2	2	1	2	2
F	68	2	1	2	1	1	2	1
M	53	2	2	2	2	2	1	2
F	61	2	2	2	2	2	2	1
M	72	1	1	1	1	2	2	2
F	60	2	1	1	1	1	2	1
M	58	2	1	1	1	1	2	2
M	69	2	1	1	1	1	1	2
F	48	1	2	2	2	2	2	2
M	75	2	1	1	1	2	1	2
M	57	2	2	2	2	2	1	1
F	68	2	2	2	2	2	2	1
F	61	1	1	1	1	2	2	1
F	44	2	2	2	2	2	2	1
F	64	1	2	2	2	1	1	2
F	21	2	1	1	1	2	2	2
M	60	2	1	1	1	1	2	2
M	72	2	2	2	2	2	1	2

Fig 1:Dataset used for model creation and evaluation

B. Dataset Visualization

Data visualization is a powerful tool for understanding complex datasets by translating them into visual forms. Through the use of charts, graphs, maps, and other graphical elements, data visualization provides a means to explore, analyze and relationships within the data.

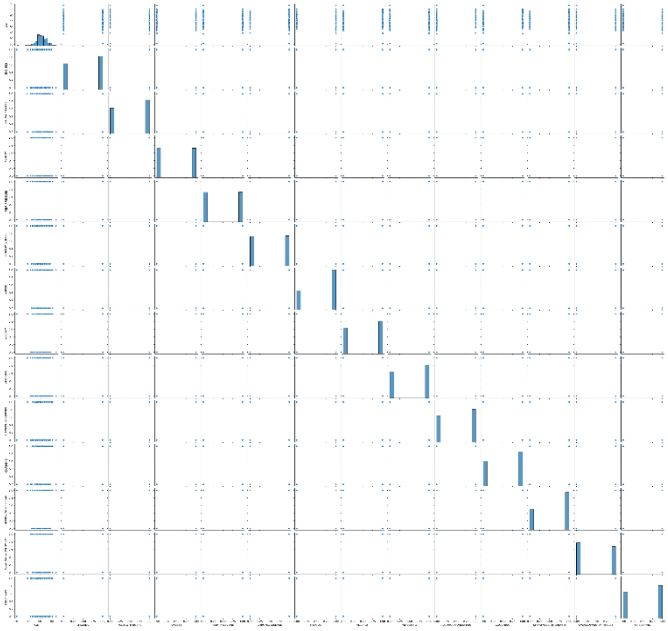

Fig.2 Data Distribution of Each Attribute

The figure 2 above plot represents how data is related from one attribute to another attribute

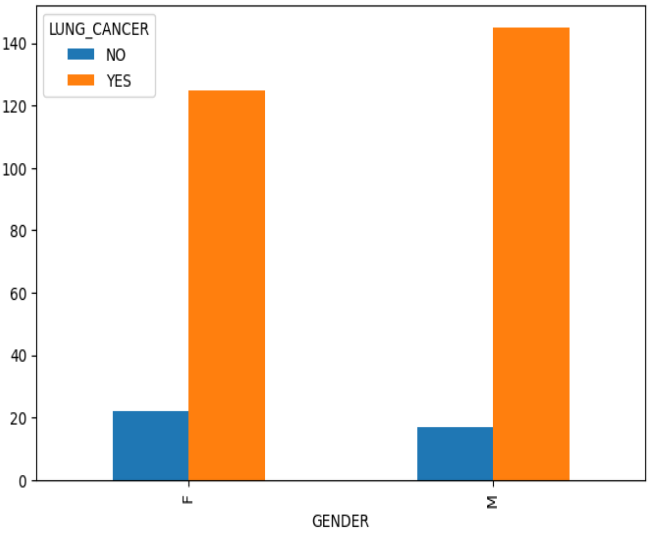

Fig 3: Total no of males and females having cancer

Figure 3 describes the total number of males and females having the cancer . The Figure represents the total count of smoking persons present in the dataset. The Figure 5 pie chart represents the count of patients who is having yellow fingers in the dataset.

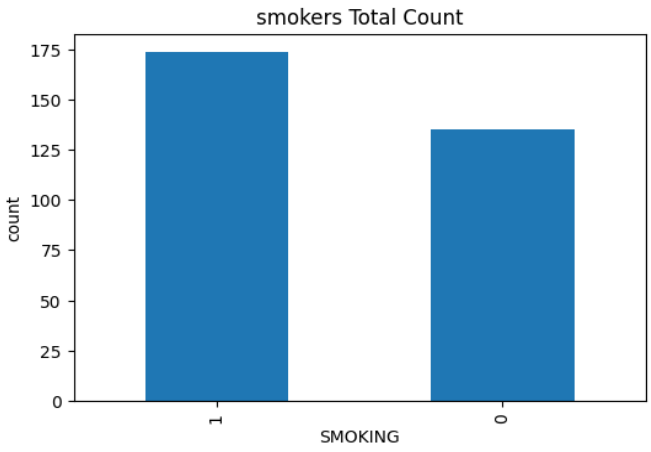

Fig:4:Total count of smoking persons in the dataset

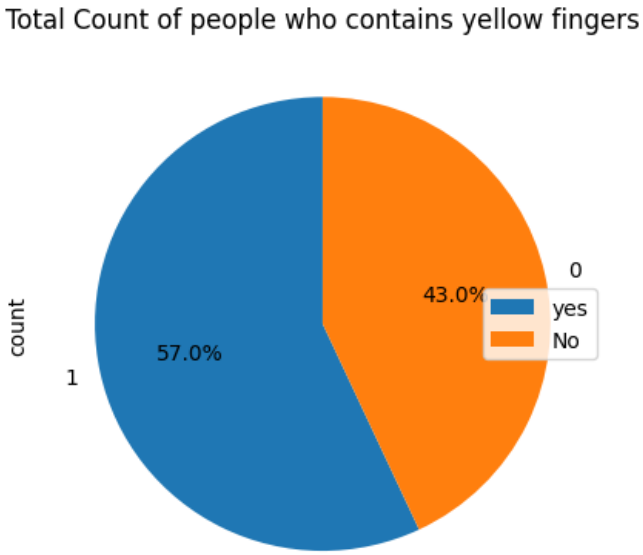

Fig 5: Count of Yellow_Fingers

C. Preprocessing Techniques

Preprocessing techniques refer to a set of procedures or methods used to prepare and clean data before it is analyzed or used for machine learning tasks. These techniques are crucial to ensure that the data is in a suitable format, free from errors, inconsistencies, and irrelevant information.

Techniques used in are removal of null values, finding correlation, detecting outliers and applying [6] SMOTE. After applying SMOTE target attribute is converted as balanced class.

The association between the attributes in the dataset is shown in Figure 6. Anxiety and yellow finger features are highly correlate after applying correlation. Next, remove the single attribute from the dataset either anxiety or yellow fingers

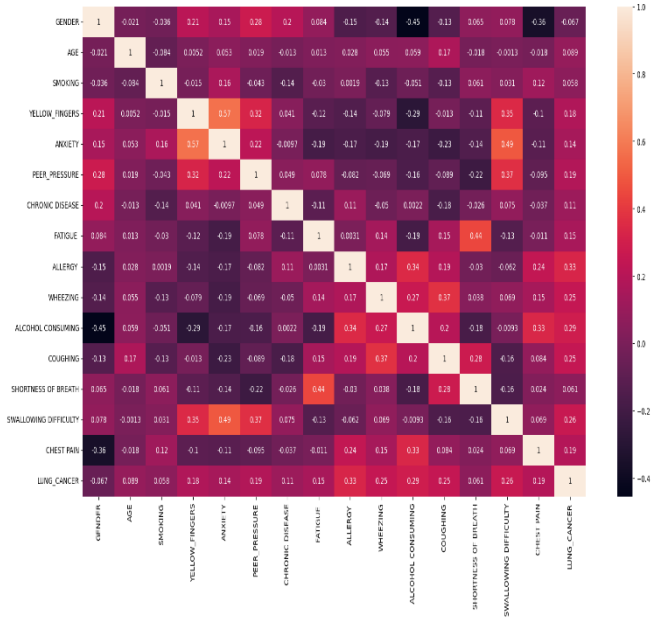

Fig 6:correlation of the entire dataset

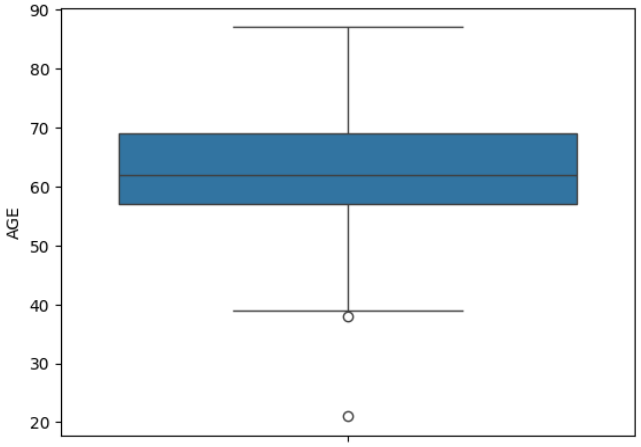

Fig:7 outliers present in age attribute

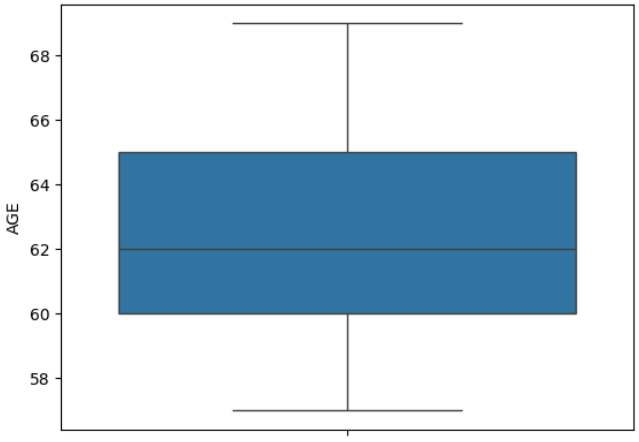

Fig:8 after removing outliers in age attribute

Figure 7 and Figure 8 represents the outliers removing in the age attribute present in the dataset.outliers can be visualized by using the box plot.

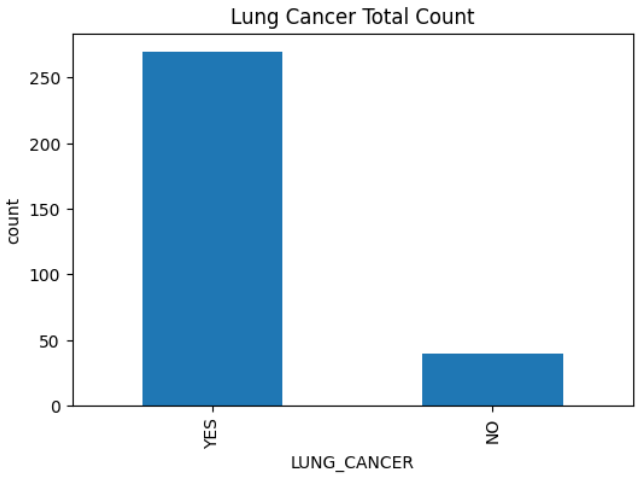

Fig:9 Total count of target attribute

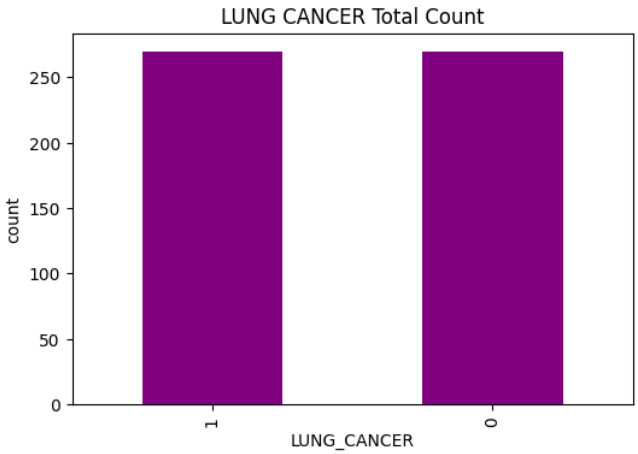

Fig :10 after applying smote algoritham

Figure 9 illustrates the total count of attributes related to lung cancer. To address the class imbalance in the target attribute, the SMOTE algorithm is employed, as depicted in Figure 10.

D. Classification of Data

Data classification involves grouping or tagging data into predefined classes or categories using their attributes or features. This process, frequently employed in machine learning and data mining, aims to construct predictive models capable of automatically assigning new data instances to suitable classes. The datasets must be tested and trained; out of the 303 datasets, 50%must be trained, and 50% must be tested.

Example code:

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(x,y, test_size = 0.2, random_state =42)
print(x.shape,X_train.shape,X_test.shape,y.shape)
```

E. Model Creation and Evaluation

1.LogisticRegression

Logistic regression, a supervised learning method, predicts the anticipated value for a dependent group using inputs like positive/negative or yes/no data[7]. It generates a stochastic value by fitting a sigmoid-shaped logistic function with maximum load possibilities of 0 and 1, reminiscent of linear regression. The linear regression's curve indicates malignant cells, and figure 4 shows the confusion matrix for the logistic regression model.

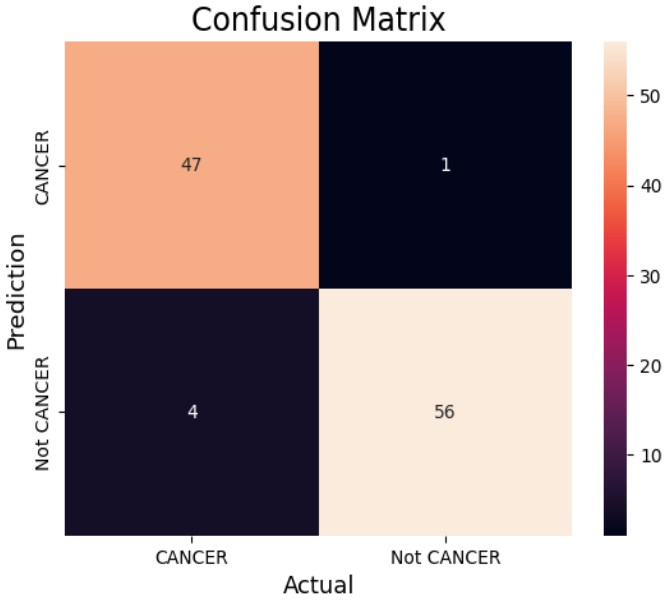

Fig.11 Confusion matrix for Logistic Regression

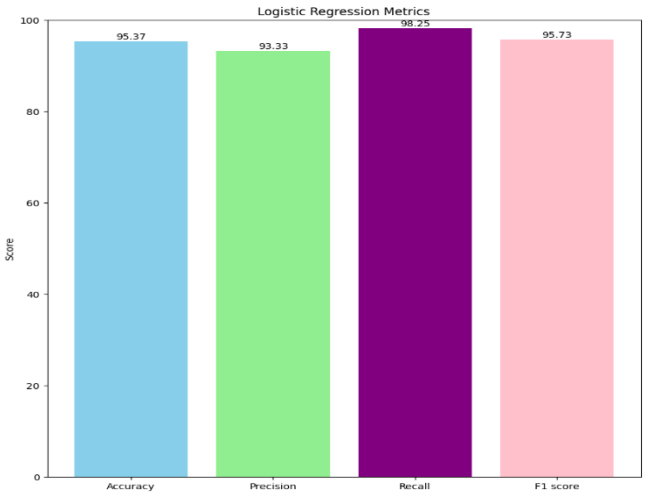

Fig:12 Before cross validation metrics of LR

Figure 12 and 13 represnts the metrics, before performing the cross validation on the model.

	precision	recall	f1-score	support
0	0.921569	0.979167	0.949495	48.000000
1	0.982456	0.933333	0.957265	60.000000
accuracy	0.953704	0.953704	0.953704	0.953704
macro avg	0.952012	0.956250	0.953380	108.000000
weighted avg	0.955395	0.953704	0.953812	108.000000

Fig:13 Classifiction report brfore cross validation

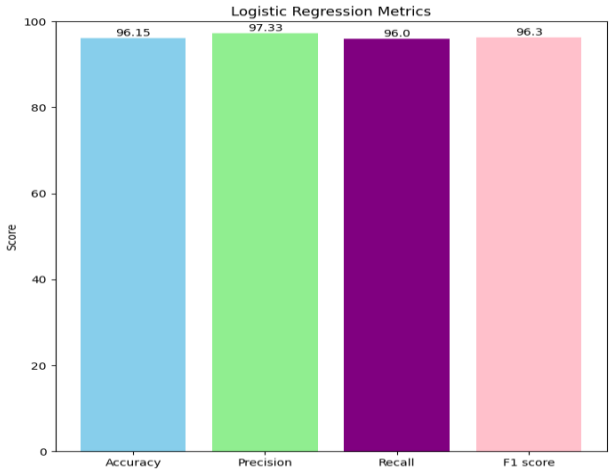

Fig:14 After cross validation metrics of LR

	precision	recall	f1-score	support
0	0.959410	0.962963	0.961183	270.000000
1	0.962825	0.959259	0.961039	270.000000
accuracy	0.961111	0.961111	0.961111	0.961111
macro avg	0.961117	0.961111	0.961111	540.000000
weighted avg	0.961117	0.961111	0.961111	540.000000

Fig:15 Classification report after cross validation

Figure 14 and 15 represnts the metrics, after performing the cross validation on the model.

2. Decision Tree

The decision tree[8] method is a popular supervised learning technique used for regression and classification challenges. It involves training the model to forecast class by learning from previous data, identifying the root class label and branching into leaf and decision nodes. Confusion matrix for this algorithm as shown in below figure 16.

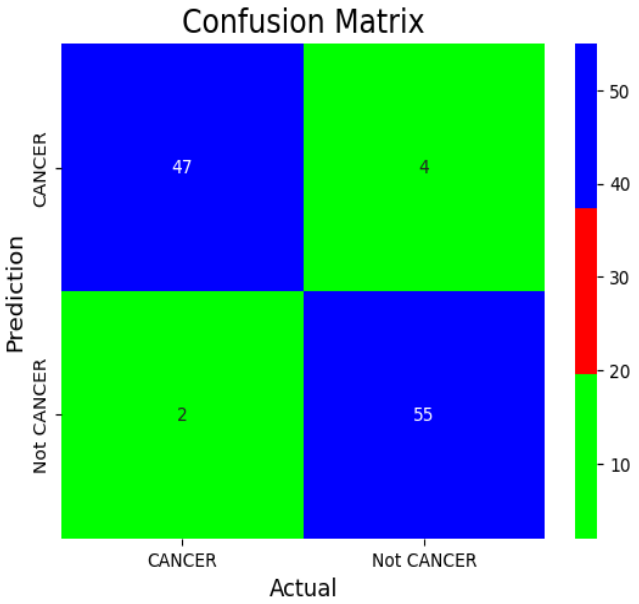

Fig 16:Confusion Matrix for Decision Tree

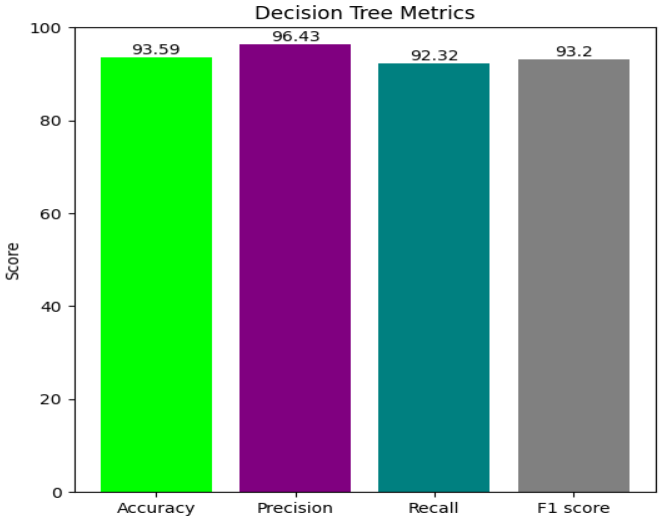

Fig:17 After cross validation metrics of DT

	precision	recall	f1-score	support
0	0.924188	0.948148	0.936015	270.000000
1	0.946768	0.922222	0.934334	270.000000
accuracy	0.935185	0.935185	0.935185	0.935185
macro avg	0.935478	0.935185	0.935174	540.000000
weighted avg	0.935478	0.935185	0.935174	540.000000

Fig:18 classification report after cross valiodation

Figure 17 and 18 represnts the metrics, after performing the cross validation on the model.

3. Random Forest

Random Forest is a popular machine learning[9] technique for classification, regression, and other tasks.. Each tree in Random Forest is trained on a random subset of features and data, improving accuracy and reducing overfitting. The below Figure 19 represents the RF algorithm confusion matrix.

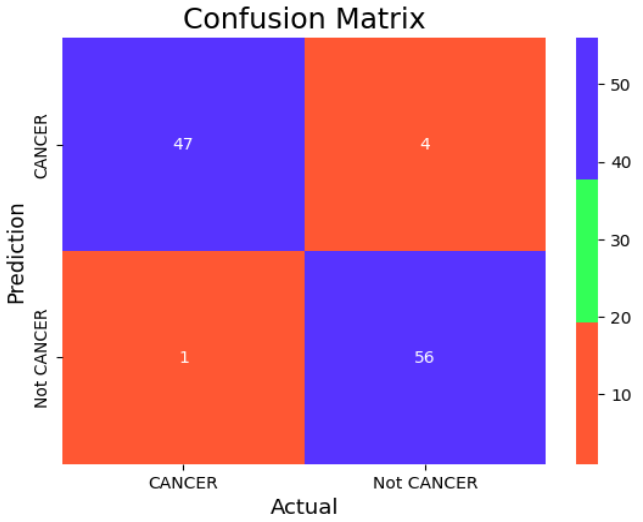

Fig.19 Confusion matrix for Random Forest

Figure 20 displays the performance metrics such as accuracy, precision, F1 score, and recall for the random forest algorithm following k-fold cross-validation. Figure 21 display the classification report after cross validation.

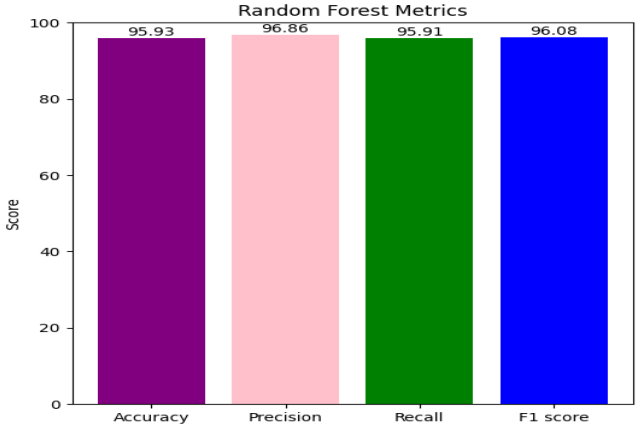

Fig:20 After cross validation metrics of RF

	precision	recall	f1-score	support
0	0.959259	0.959259	0.959259	270.000000
1	0.959259	0.959259	0.959259	270.000000
accuracy	0.959259	0.959259	0.959259	0.959259
macro avg	0.959259	0.959259	0.959259	540.000000
weighted avg	0.959259	0.959259	0.959259	540.000000

Fig:21classification report after cross validation

4.K-Nearest Neighbour

KNN is a straightforward ML algorithm based on supervised learning, using feature [10]similarity to determine new data values through the Euclidean distance of K nearest neighbours. The below figure 7 describes the confusion matrix for this algorithm.

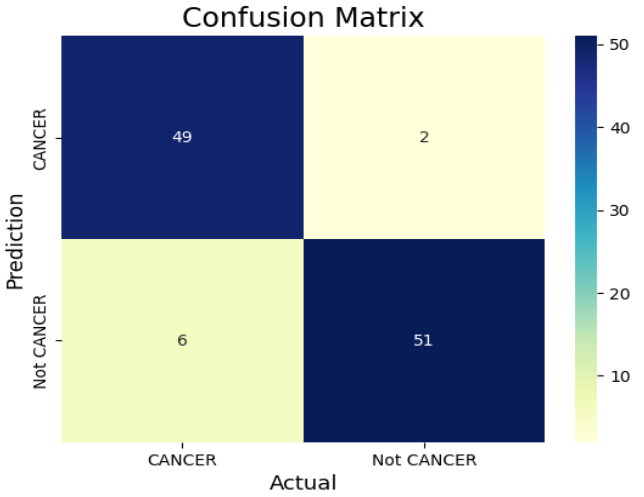

Fig 22 :confusion Matrix for KNN

Figure 23 displays the performance metrics such as accuracy, precision, F1 score, and recall for the knn algorithm following k-fold cross-validation. Figure 24 display the classification report after cross validation of KNN

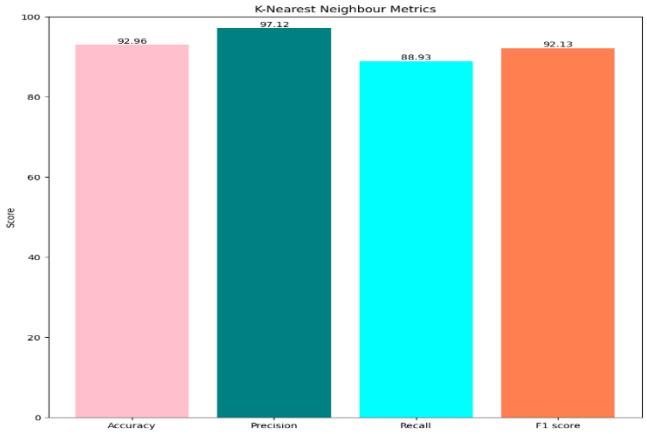

Fig:23After cross validation metrics of KNN

	precision	recall	f1-score	support
0	0.897260	0.970370	0.932384	270.00000
1	0.967742	0.888889	0.926641	270.00000
accuracy	0.929630	0.929630	0.929630	0.92963
macro avg	0.932501	0.929630	0.929513	540.00000
weighted avg	0.932501	0.929630	0.929513	540.00000

Fig:24 classification report after cross validation

5.Support Vector Machine

SVM is a popular ML technique for regression and classification[11], creating a hyperplane . It optimizes the margin between classes to find the best hyperplane[12] .Below figure 25 describes the confusion matrix for this algorithm.

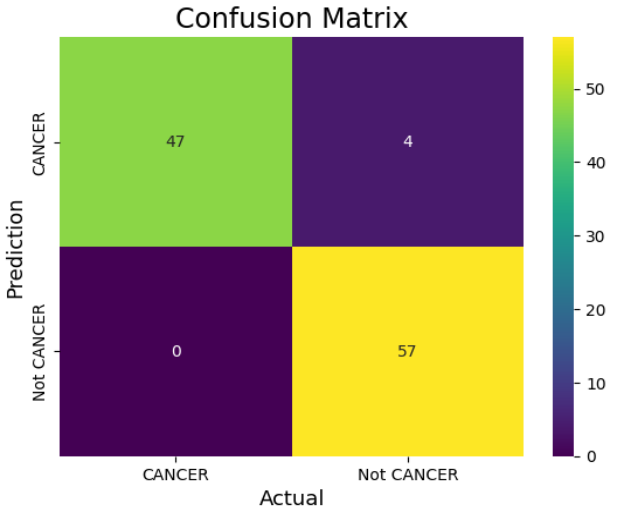

Fig:25 Confusion matrix for Support Vector Machine

Figure 26 displays the performance metrics such as accuracy, precision, F1 score, and recall for support vector machine algorithm following k-fold cross-validation. Figure 24 display the classification report after cross validation of SVM

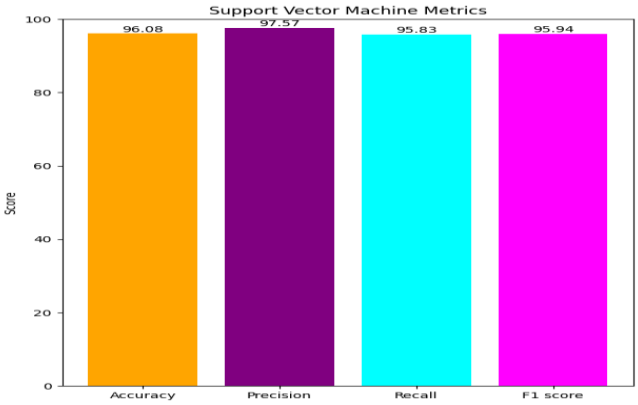

Fig:26 After cross validation metrics of SVM

	precision	recall	f1-score	support
0	0.955882	0.962963	0.959410	270.000000
1	0.962687	0.955556	0.959108	270.000000
accuracy	0.959259	0.959259	0.959259	0.959259
macro avg	0.959284	0.959259	0.959259	540.000000
weighted avg	0.959284	0.959259	0.959259	540.000000

Fig:27 classification report after cross validation

6.Naive Bayes

The basic Bayes algorithm, a supervised learning strategy applying Bayes theorem to classification [13], is a straightforward method for machine learning model creation. It uses a probabilistic classifier to predict item likelihood based on data. Figure 9 represents the confusion matrix for this algorithm.

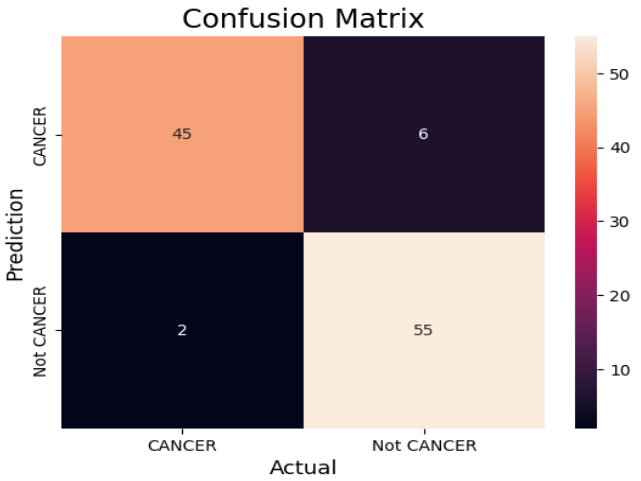

Fig :28 Confusion Matrix for Naïve Bayes

Figure 23 displays the performance metrics such as accuracy, precision, F1 score, and recall for the decision tree algorithm following k-fold cross-validation. . Figure 24 display the classification report after cross validation of NB

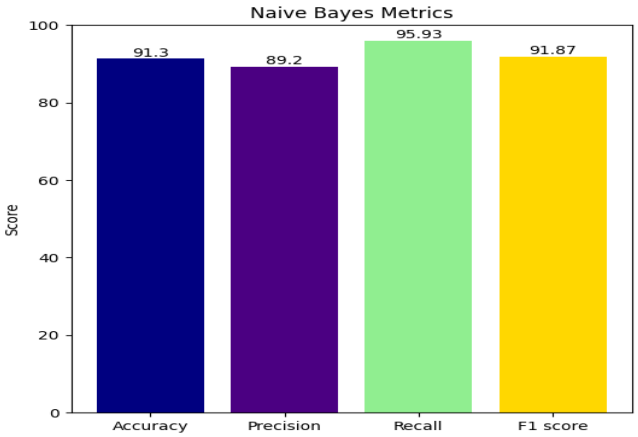

Fig 29*:* After cross validation metrics of NB

	precision	recall	f1-score	support
0	0.955102	0.866667	0.908738	270.000000
1	0.877966	0.959259	0.916814	270.000000
accuracy	0.912963	0.912963	0.912963	0.912963
macro avg	0.916534	0.912963	0.912776	540.000000
weighted avg	0.916534	0.912963	0.912776	540.000000

Fig:30 classification report after cross validation

IV. Result and Analysis

In this study, several machine learning algorithms used for early detection of lung failure, as described in the ml algorithm[14] above in this paper. Therefore, using the records in the dataset provides the highest accuracy[15] using the logistic algorithm shown in below Table 1.

Table−1 Comparison of all models

Model	Accuracy	Precision	Recall	F1_score
Logistic Regression	96.14	97.33	96.00	96.29
Decision Tree	93.58	96.42	92.32	93.20
Random Forest	95.92	96.86	95.91	96.08
Support Vector Machine	96.08	97.57	95.83	95.94
K-Nearest Neighbour	92.96	97.12	88.93	92.13
Naïve Bayes	91.29	89.19	95.93	91.87

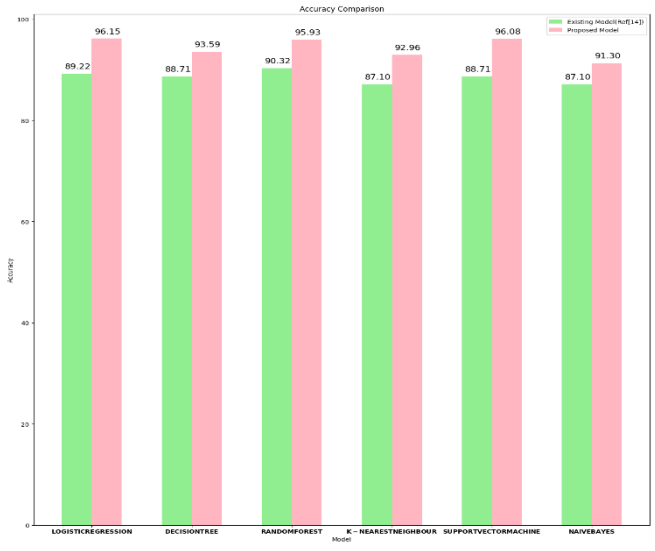

Fig.31 comparison of previous model and existing model

Figure 31 describes the accuracy values of previous model and Existing model. Figure 32, Figure 33,Figure 34,Figure 235 describes the accuracy ,precision f1score,recall for all models present in the existing system.

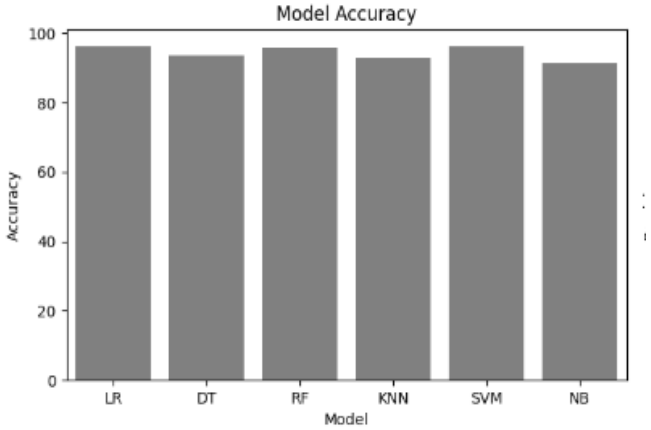

Fig:32 comparison of all models accuracy

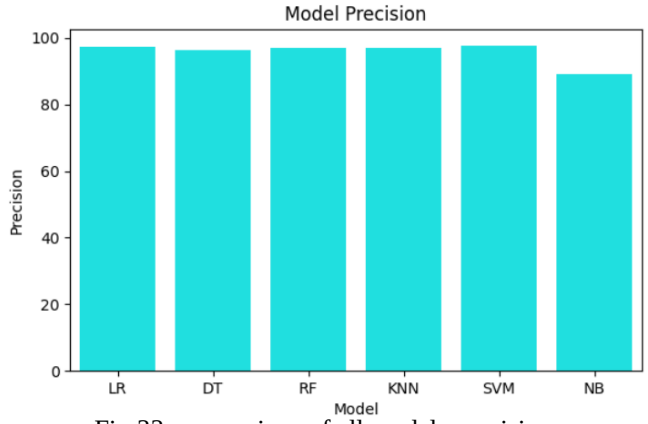

Fig:33 comparison of all models precision

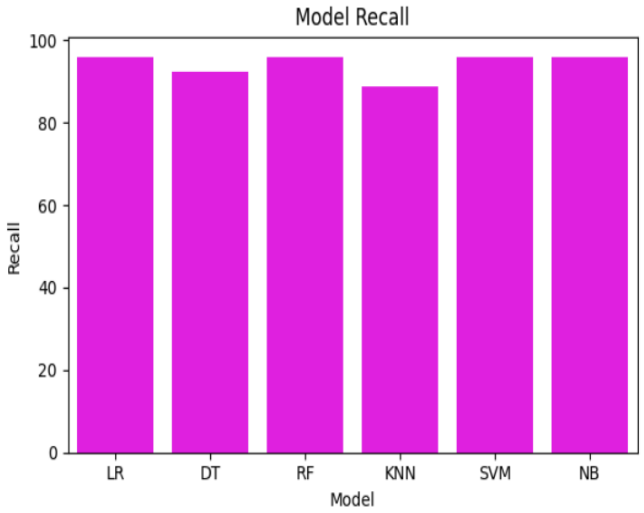

Fig:34 comparison of all models recall

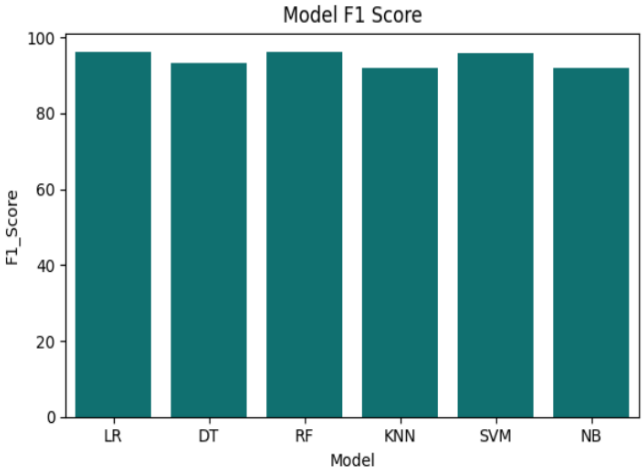

Fig:35 comparison of all models f1score

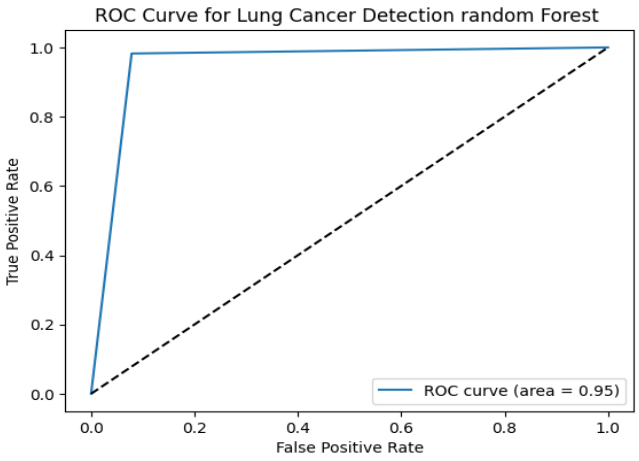

Fig . 38 Roc Curve for RF

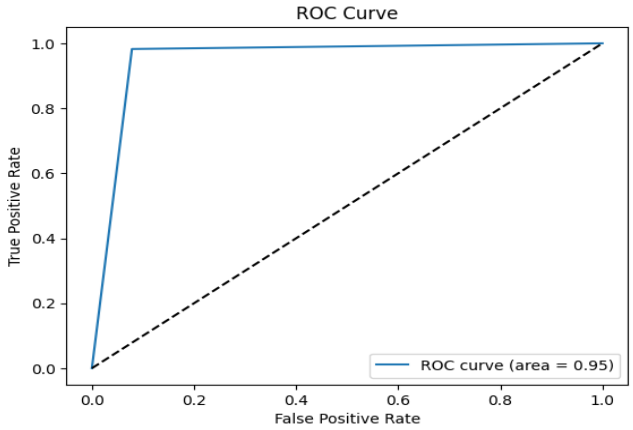

Fig:36 Roc Curve for Logistic regression

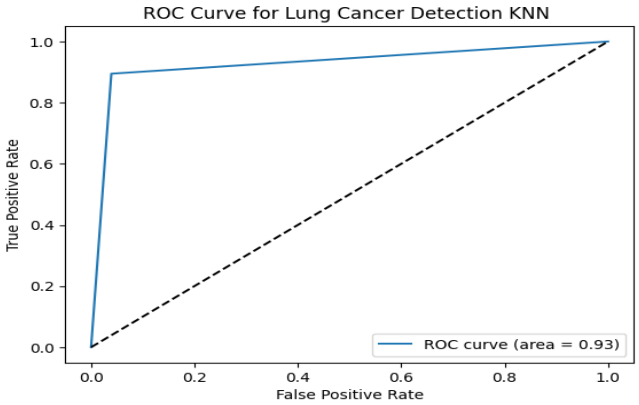

Fig . 39 Roc Curve for KNN

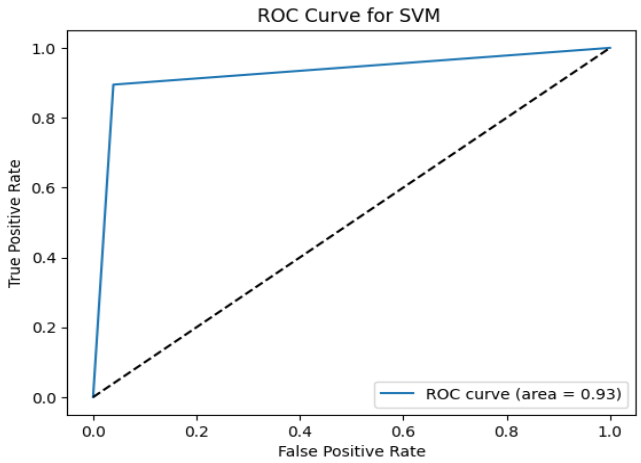

Fig :37 Roc Curve for SVM

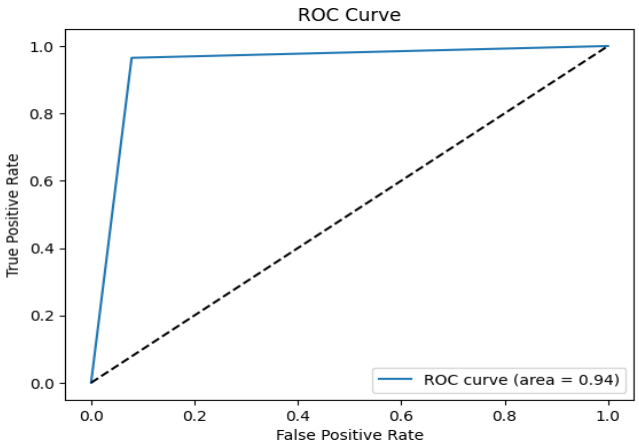

Fig . 40 Roc Curve for DT

Figure36,Figure37,Figure38.Figure 39,Figure 40 represents the roc curve of LR,DT,KNN,SVM,RFalgorithms

V..Conclusion and Futurescope

We have used algorithms like Logistic Regression provides accuracy of 96.1%, Decision Tree have provides accuracy of 93.5%, Random Forest have provides accuracy of 95.9%. K-Nearest Neighborshave provides accuracy of 92.9%, naïve bayes have provides accuracy of 88.17%,SVM have provides accuracy of 96.0%.After comparing experimental results, we have found that LR provides the highest accuracy of 96.1% Regression is more when K-Fold cross validation is applied. In future studies, we recommend addressing other topics such as feature selection process techniques, as well as trying explore different approaches to classification models, including machine learning methods and improved interpretation of regression models.

REFERENCES

[1] Rehman, M. Kashif, I. Abunadi and N. Ayesha, "Lung Cancer Detection and Classification from Chest CT Scans Using Machine Learning Techniques," 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA), 2 pp.101104,2021doi:10.1109/CAIDA51941.2021.9425269.

[2] Ajay Aggarwal , Grant Lewison , etal. The State ofLung Cancer Research:A Global Analisis" by International Association for the Study of Child Language (IASCL)published on 2with doi:10.1016/j.jtho.2016.03.010

[3] C. Anil Kumar,S. Harish,1Prabha Ravi,Murthy SVN,B. P. Pradeep Kumar,V. Mohanavel,Nouf M. Alyami,S. ShanmugaPriya"Lung Cancer prediction for textual dataset using MachineLearning". Published on 14 July 2022 DOI: doi.10.1155/2022/6254177

[4] Vikul J.Pawar,Kailash, D.Kharat,Suraj, R.Pardeshi ,Prashant D.Pathak."Lung cancer Detection using Image Processing and machine Learning Techniques"2020 by internation journal of Advanced Trends in computer science and Engineering(IJATCSE)doi:10.30534/ijatcse2020/260942020

[5] https://www.kaggle.com/code/ibrahimkaratas/lung-cancerpredictionusingmachinelearning?scriptVersionId=117035415&cellId=11

[6] BharathyS,PavithraR,akshayaB"Lung cancer Detection Using Machine Larning" ,published in 2022,International Conference on Applied Artificial intelligence and Computing(ICAAIC)on11 may 2022 ,doi:10.1109/ICAAIC53929.2022.9793061

[7] Raghavendra patil G,SinchanaC,Tejashwini P,Tejaswini K,veena Vittal Ganiga"Lung Cancer prediction using logistic Regerssion Approach" 2020 International Researuch Journal of Modernization in Engineering Technologr and Science(IRJMETS)

[8] Vamsi krishna Reddy Munnangai,"Lung Cancer Detection",by 2021 International Researuch Journal of Modernization in Engineering Technologr and Science(IRJMETS)Volume:03/Issue:11/November-2021.

[9] R. D. Karthikeyan, R. G, V. V, G. B. C and K. M, "A Review of Lung Cancer Detection using Image Processing," 2021 Smart Technologies, Communication and Robotics (STCR), pp. 1- 4, 2021,doi10.1109/STCR51658.2021.9588835.

[10] DR.SamuelManoharan,Prof.Sathish" Improved version of graph -cut Algoritham for Ct Images of lung cancer With clinical property condition "2020 by Journal of Artificial Intelligence andcapsuleNetwork(JAICN),DOI:https://doi.org/10.36548/jaicn.2020.4.002.

[11] Q. Firduas, R. Sigit, T. Harsono and A. Anwar, "Lung Cancer Detection Based On CT -Scan Images With Detection Features Using Gray Level Co-Occurrence Matrix (GLCM) and Support Vector Machine(SVM)Methods,"InternationalElectronicsSymposium(IES),2020,pp.643648, doi:10.1109/ies50839.2020.9231663.

[12] N. S. Nadkarni and S. Borkar, "Detection of lung cancer in CT Images using Image Processing," 2020 3rd International Conference on Trends in Electronics and Informatics (ICOEI), 2020, doi:10.1109/ICOEI.2019.8862577.

[13] Kumar Mohan and Bhraguram Thayyil "Machine Learning Techniques for Lung Cancer Risk Prediction using Text Dataset", International Journal of Data Informatics and Intelligent Computing pp:47-56,2023,doi:10.59461/ijdiic.v2i3.73.

[14] Kumar Mohan and Bhraguram Thayyil "Machine Learning Techniques for Lung Cancer Risk Prediction using Text Dataset", International Journal of Data Informatics and Intelligent Computing pp:47-56,2023,doi:10.59461/ijdiic.v2i3.73.

[15] Pragya Chaturvedi,Anuj Jhamb et al,"Prediction and classification of Lung C?ancer Using Machine Learning Techniques"byIOPConferenceSeries,pp:012059,2021,doi:10.1088/1757899X/1099/1/012059