Performance Evaluation of Different Machine Learning Algorithms for the Detection of Lung Cancer

M. Prema Kumar[1], V. Veer Raju[2], M. Venkata Subbarao[3], G. Challa Ram[4]
[1,3,4]Dept. of ECE, Shri Vishnu Engineering College for Women (A), Bhimavaram, A.P.
[2] Dept. of ECE, Sir C.R.R.C.o.E. Eluru, A.P.

Abstract.

Cancer is the main reason for a huge number of deaths worldwide, out of which lung cancer is the main cause of the highest mortality rates. Nearly 85% of males and 75% of females suffer from lung cancer. The cancer cells grow and keep multiplying which leads to the development of tumors. If these cells grow rapidly, it spreads to other parts of the body, this is known as Metastases. Identification of cancer at the final stage has very less chances of complete treatment and it might lead to the death of the patient. Therefore, early prediction before the final stage is highly essential to increase the survival rate. For early identification, various Machine learning techniques are used which will facilitate the fast treatment of the disease. The dataset consists of different attributes such as Smoking, Alcohol consumption, Chest pain, Shortness of Breath, etc. The various ML classifiers applied to the dataset are Decision tree, Logistic regression, SVM, Naive Bayes, KNN, and Random forest. The classification models are analyzed for different test and train ratios and the obtained Accuracy, Precision, Recall, Error rate, Specificity, F-Measure, and Time are noted. This process is carried out for both binary and multi-class classification. Multiclass classification considered here is 3 class classification, i.e. High, Low and Medium levels of lung cancer. The accuracy obtained tells up to what extent the classifier has correctly predicted the disease. The highest accurate model for both the classes obtained was SVM Classifier. The accuracy obtained was 100% with minimum execution time. Therefore, the SVM classifier using the Machine Learning technique can be applied to detect the presence of the disease and hence help the doctors in identifying it. By doing so, early diagnosis can be performed and required precautions can be taken.

Keywords: Lung Cancer, SVM, Machine Learning, Metastases.

DOI Number: 10.14704/NQ.2022.20.12.NQ77262 **NeuroQuantology2022;20(12): 2692-2699**

1. Introduction

The Human body is a combination of tiny particles called cells, these cells will grow unconditionally when cancer is present in the human body. These cells will grow to form a tumor. If cells are there in the body for longer period, they can even spread to other body parts. This is known as metastasis. Cells that were damaged in lungs cause Lung cancer. Other possible cancer types, such as kidney or breast, can metastasize to the lungs. When it was happened, this is no longer lung cancer. For Ex. if breast cancer extents to the lungs, it will be considered as metastatic breast cancer. [1]
Lung Cancer – A Deadly disease

The lungs feel like sponge type tissues in the chest. Its job is to pass oxygen (O_2) to the body and leave the carbon dioxide (CO_2). When human body breathes air, it drives into lungs via windpipe (trachea). The trachea splits into pipes called bronchi, which goes to the lungs. These divided tiny branches called bronchioles. At the end its tiny air sacs called alveoli. The alveoli transfer oxygen from the air into the blood. It takes CO_2 out of the blood and leaves human body when breathe out (exhale).
Lung cancer is the second highest among men and women with roughly 2.09 million new detects every year, and it is accountable for the more deaths, at nearby 1.76 million, w.r.t the World Health Organization (WHO).

2. Related Work

Janee Alam, Sabrina Alam and Alamgir Hossan et. al proposed "Multi-Stage Lung Cancer Detection and Prediction Using Multi-class SVM (Support Vector Machine) classifier". This is used to detect the cancer and could predict the probability of occurrence. In this image enhancement along with segmentation have done distinctly. Some Image transformation techniques have been used in enhancement, coming to the segmentation Thresholding and marker-controlled watershed based segmentation has been used. SVM Classification binary classifier was used and it can detect affected cell and its phase such as initial, middle, or final stage. If no cell was affected, then it looks for the probability of occurrence of lung cancer [1].

Qing Wu and Wenbing Zhao et al. proposed "Small-Cell Lung Cancer Detection Using a Supervised Machine Learning Algorithm". It's a novel neural-network based algorithm, and it refers to an Entropy Degradation Method (EDM), to identify small cell lung cancer (SCLC) from CT Scan images. It assists premature detection of lung cancers. The training and testing data utilizes high resolution

eISSN 1303-5150

lung CT scan images. They selected 12 lung CT scan images, 6 are from healthy lungs and 6 are from patients with SCLC and to train the model take randomly 5 scans from each set, and used two scans to test and it gave an accuracy of 77.8% [2].

3 Dataset Description

Dataset: Data is an integral part of the algorithm design, it is important to have a clean and correctly labelled dataset. By inputting accurate data into the algorithm we will have accurate outputs, resulting in a more effective and timely training process. The dataset for 2 class classification consists of 310 member's data with 15 attributes. The attributes given are: Gender, Sex, Coughing, Shortness of breath, Anxiety, Smoking, Alcohol, Allergy, Peer Pressure, Chronic Disease, Yellow Fingers, Fatigue, Wheezing, swallowing difficulty and Chest pain. For first analysis all the 15 attributes are considered. For the second analysis 3 attributes Gender, Age and Allergy were removed. Similarly, for the third analysis 2 more attributes yellow fingers and peer pressure were removed.

The above dataset was collected from Kaggle. It permits users to find and publish data sets, build and explore models and work with other data and ML scientists, and enter competitions to crack data science tests.

Training data is the sub set of the data and is used to assistance the Machine Learning (ML) model to create predictions and run the model for data thoroughly. Testing data arises into picture after a lot of development and validation. Probing the model to mark predictions based on the data is meant to test whether it will work or not in real time.

4 Classification Methods

Classification is from supervised learning to predict given data to a certain class label. The uniqueness in classification depend on mapping function to a firm output level. Numerous learning classifiers were labelled as Perceptron, Naïve Bayes, Decision Tree, Logistic Regression, K Nearest Neighbor, Artificial Network, Support Vector Machine(SVM). Classification in ML is one of earlier decision method used for data analysis. The idea of the work focuses on novel approach of ML for analysis of lung cancer data to attain a decent accuracy. Some of the mostly used classifiers are described as Decision tree, Logistic Regression, Naive Bayes, SVM, KNN and Random Forest. After testing and training the data, various parameters like Accuracy, Precision, Recall, Error rate and Specificity are calculated.

The ML algorithms are further classified as:

4.1. Decision Tree:

Classification involves two steps learning and prediction. In this learning, the model is developed based on certain training data. In the prediction, the model is used to predict the answer for given data. It is the easiest and prevalent classification algorithm to understand and interpret the data. It belongs to supervised learning [3]. It is used for solving regression and classification problems. The main aim of this is to produce a training model that which can be used to predict the class.

4.2. Logistic Regression

It is classified into three categories:

- •Binomial will have two likely types of the dependent variables, such as 1 or 0, Pass or Fail, etc.
- •Multinomial will have three or more likely unordered types of the dependent variable, such as cats, dogs, or sheep.
- •Ordinal will have three or more likely ordered types of dependent variables, such as low, Medium, or High.

4.3. Naive Bayes:

Bayes theorem used for solving problems related to classification. It is primarily used in *text classification* that includes a high-dimensional training data. It is simple and utmost active Classification algorithms which helps in quick predictions based on the fast ML models. It is a probabilistic classifier; to predict the basis of the probability of an object**.**

4.4. SVM

It is a popular Supervised Learning algorithm and used for Classification and Regression problems. The main objective of the SVM is to produce the best line which can separate N-dimensional space into modules, hence it can easily have kept the new data in the correct category in future. This supreme decision boundary is called a hyper plane. SVM indicates the extreme points/vectors that helps in generating the hyper plane. The great points are called as support vectors, and hence it is labeled as SVM.

4.5. KNN

- Fine KNN
- Medium KNN

K-NN is the simplest Supervised Learning. K-NN algorithm accepts and stores the similar data and available data and placed the new into the category that is most alike to the obtainable categories. It means when new data give the idea then it can be easily classified into a fine set category by using K- NN algorithm.

4.6. Random Forest Algorithm

It is a prevalent ML algorithm comes under supervised learning. Classification and Regression can be done using this. It is based on ensemble learning and it uses multiple classifiers to resolve complex problems and to progress its performance.

5. Proposed Model

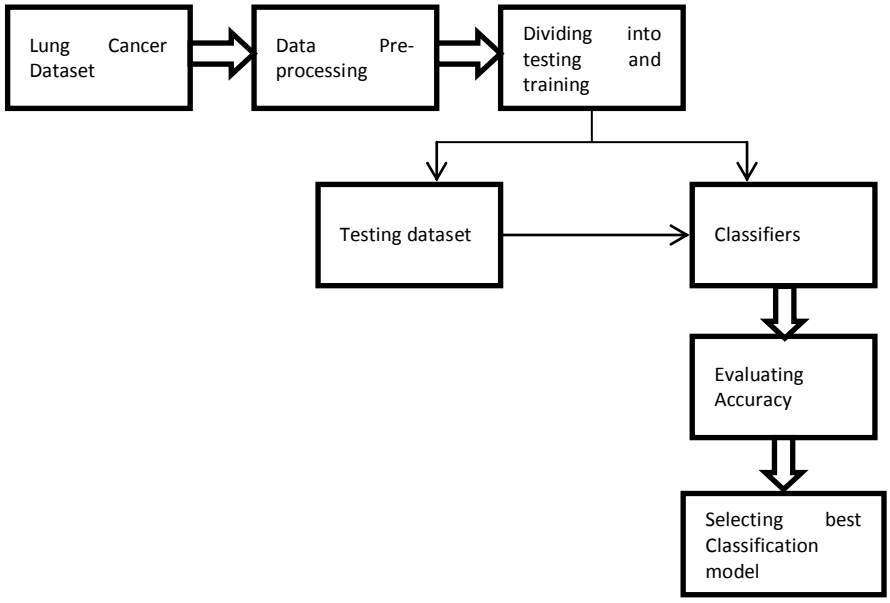

Fig 1: Block diagram

The above shown is the block diagram of proposed work. It describes the process involved in the implementation of proposed work.

- Firstly, the lung cancer dataset is gathered and pre-processing is performed.
- The data is trained under different ML classifiers.
- Accuracy for different test and train ratios is observed
- Comparison of all the classifiers is done.
- Attributes that do not cause any change to accuracy are removed.
- Again train the new data with different test and train ratios for different classifiers.
- Compare the performances of obtained new and previous data.
- Repeat the steps until we get more accuracy with les number of attributes.

eISSN 1303-5150

5.2 Flow Chart:

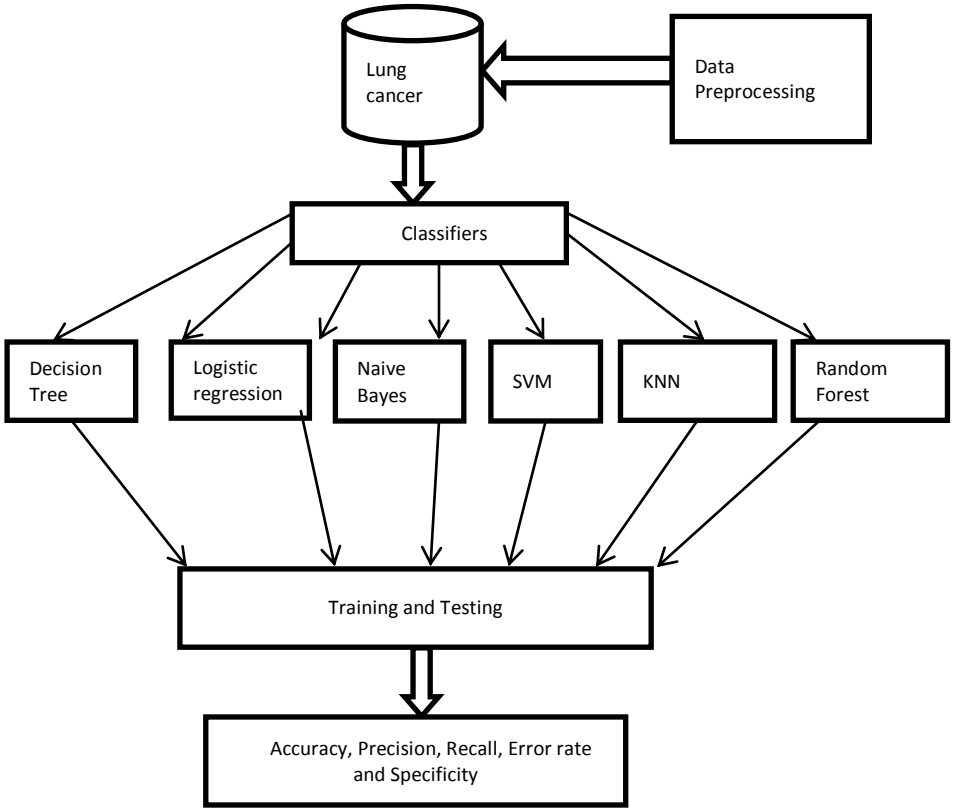

Fig 2: Flow chart

6. Result Analysis

In general confusion matrix is having key parameters like Accuracy, Recall, Precision and F-Measure for classification. Accuracy is the correct measure of predictions made out of total predictions. These Quantitative parameters depends on specific outcome. Those are True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN).

The calculations are given below:

Accuracy: It defines model correct predictions. It's the ratio of the Total no. of correct predictions made by the classifier to all the Total no. of predictions made by the classifiers. The formula is as given below:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad Eq.\ (1)$$

Precision: It defines the no. of correct outputs provided by the model, how many of them were actually true. The formula is as given below:

$$Precision = \frac{TP}{TP+FP} \quad Eq.\ (2)$$

Recall: It defines total positive classes predicted correctly. It must be as high as possible. The formula is as given below:

$$Recall = \frac{TP}{TP+FN} \quad Eq.\ (3)$$

Error Rate: It tells how frequently the classifier is wrong.

$$Error\ Rate = \frac{FP+FN}{TP+TN+FP+FN} \quad Eq.\ (4)$$

eISSN 1303-5150

Specificity: It tells how well a test can identify true negatives.

$$Specificity = \frac{TN}{TN+FP} \quad Eq.\ (5)$$

F-measure: It is a measure of the accuracy of the test.

$$F\_Measure\ \frac{2 * Recall * Precision}{Recall+Precision}\ Eq.\ (6)$$

6.1 Analysis of 2 class classification

a) Analysis with 15 attributes

Binary classification or 2 class classification refers to have two class labels. In our work the two classes are Yes or No. Yes, represents a person with lung cancer and No represents person without lung cancer. The 15 attribute considered are Gender, Sex, Smoking, Coughing, Alcohol, Allergy, Chronic Disease, Yellow Fingers, Anxiety, Shortness of breath, Peer Pressure, Fatigue, Wheezing, Swallowing difficulty and Chest pain.

The below is the confusion matrix and the result table.

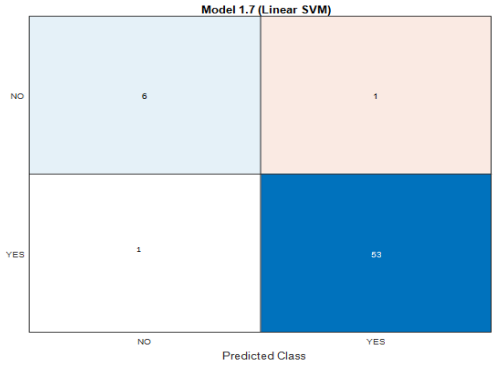

Fig 3: Confusion matrix analysis with 15 attributes

Table 1: Results for 15 attributes

Model Type	Accuracy	Precision	Recall	Error rate	Specificity	F-Measure	Time (s)
Fine Tree	93.3	100	93.1	6.66	0.33	96.42	2.1
Logistic Regression	96.7	98.1	98.1	3.2	0.8	98.1	0.8
Naive Bayes	93.4	96.3	96.2	6.5	0.7	96.2	0.8
SVM	97	98.1	98.1	3.2	0.8	98.1	0.6

b) Analysis with 12 attributes

The 12 attributes considered are Smoking, Coughing, Alcohol, Anxiety, Yellow Fingers, Peer Pressure, Chronic Disease, Fatigue, Shortness of breath, Swallowing difficulty, Wheezing and Chest pain.

The below is the confusion matrix and the result table.

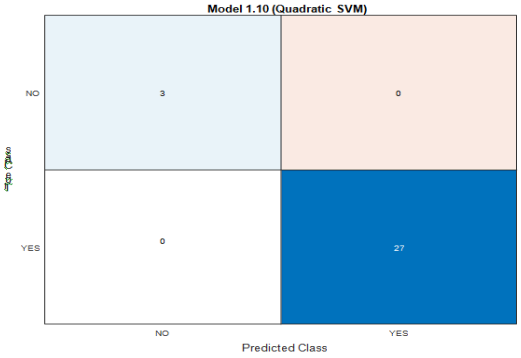

Fig 4: Confusion matrix analysis with 12 attributes

Table 2: Results for 12 attributes

Model Type	Accuracy	Precision	Recall	Error rate	Specificity	F-Measure	Time (s)
Fine Tree	96.6	96.3	100	3.3	1	98.1	7.06
Logistic Regression	100	100	100	0	1	100	1.31
Naive Bayes	93.5	96.4	96.4	6.4	0.6	96.4	1.01
SVM	100	100	100	0	1	100	0.78

eISSN 1303-5150

c) Analysis with 10 attributes

The 12 attributes considered are Smoking, Anxiety, Coughing, Alcohol, Chronic Disease, Swallowing difficulty, Shortness of breath, Fatigue, Wheezing, and Chest pain. The below is the confusion matrix and the result table.

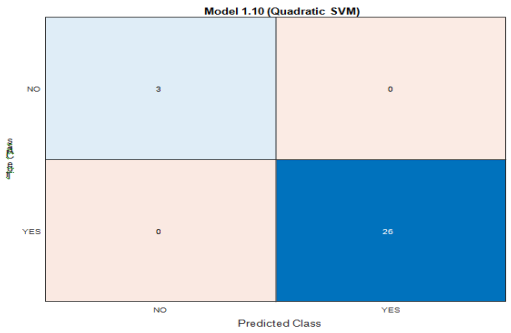

Fig 5: Confusion matrix analysis with 10 features

Table 3: Results for 10 attributes

Model Type	Accuracy	Precision	Recall	Error rate	Specificity	F-Measure	Time (s)
Fine Tree	90	92.5	96.1	10	0.6	94.3	1.2
Logistic Regression	99.3	100	99.2	0.6	0.9	99.6	23.7
Naive Bayes	94.5	97.5	96.3	5.4	0.7	96.9	5.4
SVM	100	100	100	0	1	100	8.07

Comparative analysis of calculated parameters in Machine Learning

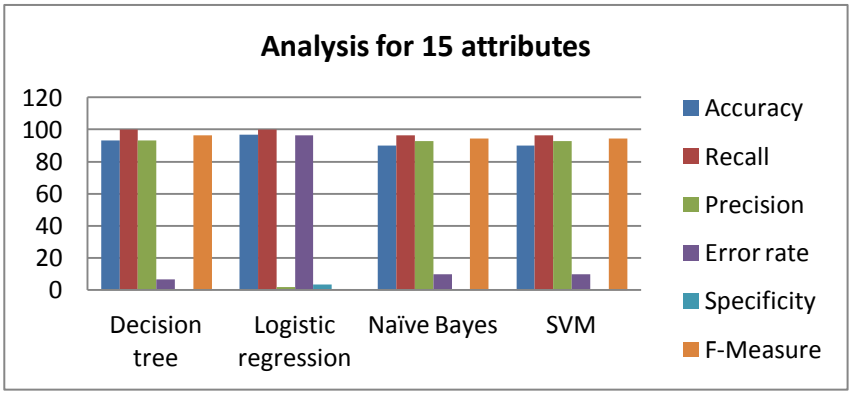

Fig 6: Analysis for 15 attributes

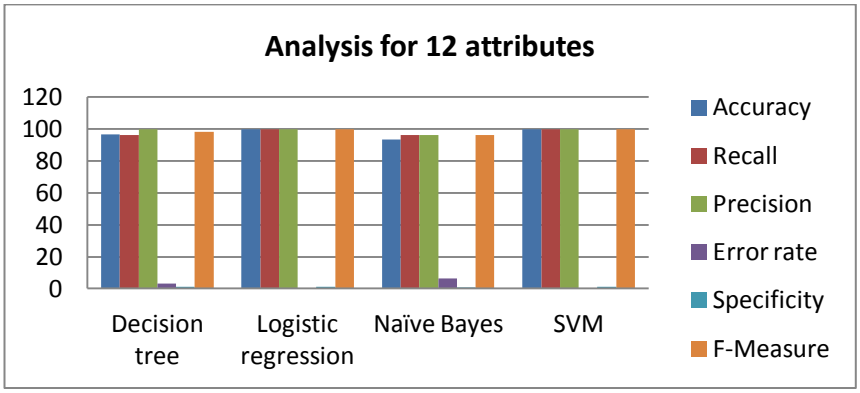

Fig 7: Analysis for 12 attributes

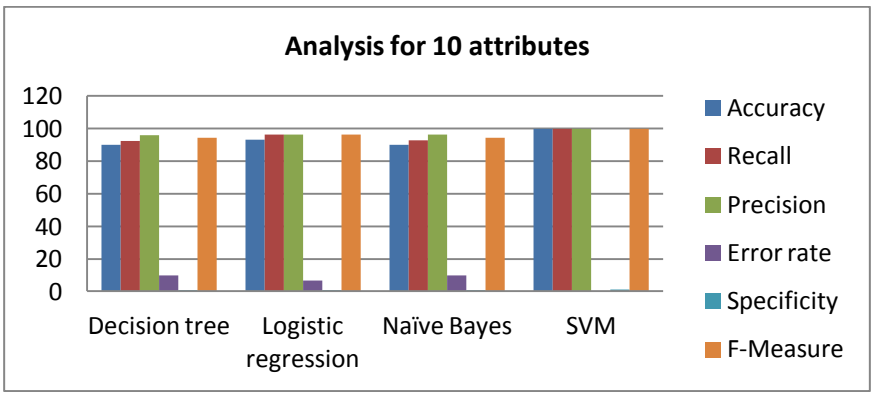

Fig 8: Analysis for 10 attributes

6.2a. Graphical analysis:

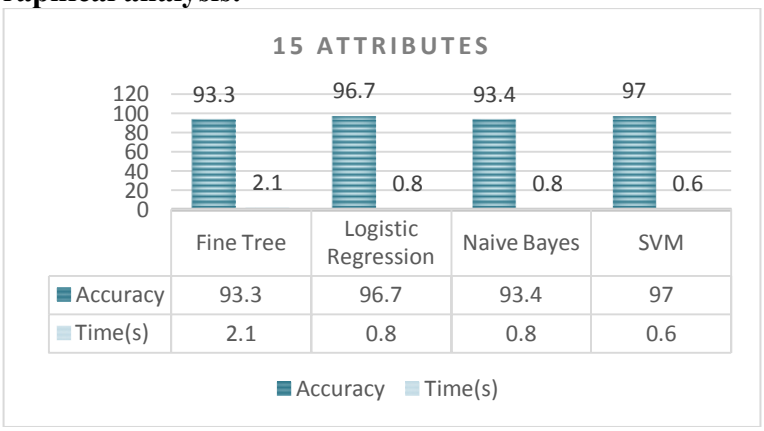

Fig 9: Graph for 15 attributes

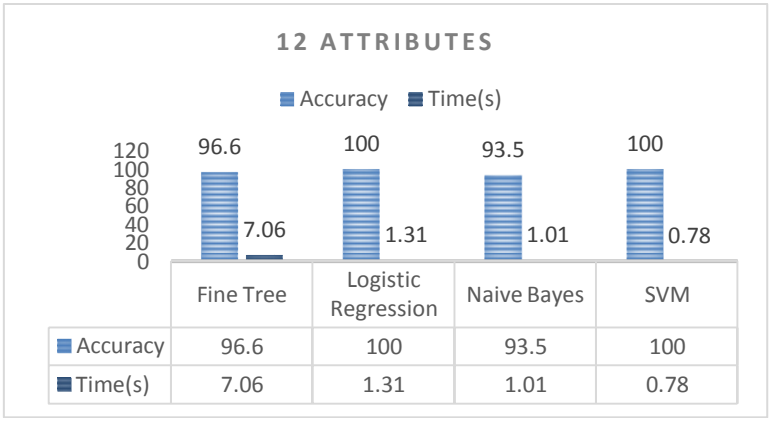

Fig 10: Graph for 12 attributes

eISSN 1303-5150

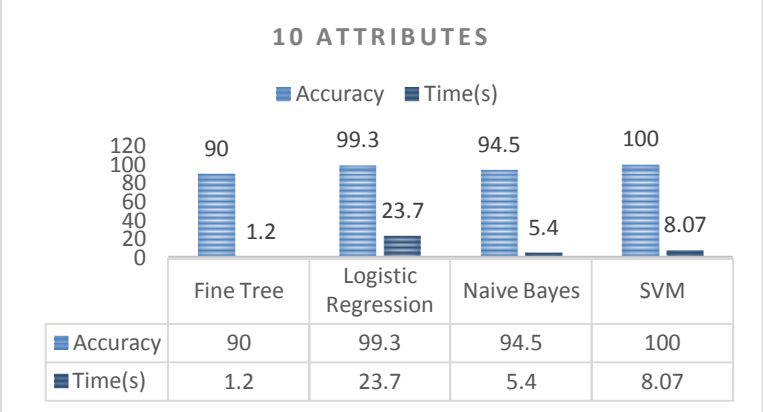

Fig 11: Graph for 10 attributes

From the various sub classifications present for decision tree, Naïve Bayes and SVM -- Fine tree, Gaussian Naïve Bayes and Linear SVM have been considered as it has given better accuracy than the other methods.

7. Conclusion

From the results obtained, for 2 class classification and 15 attributes taken we have obtained an accuracy of 93% for fine tree, 96.7% for Logistic regression, 93.4% for Naïve Bayes, and 96.7% for SVM. With 12 attributes taken, we have obtained an accuracy of 96.6% for fine tree, 100% for logistic regression, 93.5% for Naïve Bayes and 100% for SVM. With 10 attributes taken, we have obtained an accuracy of 90% for fine tree, 99.3% for logistic regression, 94.5% for Naïve Bayes and 100% for SVM.

References

[1] Janee Alam, Sabrina Alam, Alamgir Hossan, "Multi-Stage Lung cancer detection and prediction using multi-class SVM classifier", International Journal of Innovative Technology and Exploring Engineering ,2019.

[2] Qing Wu and Wenbing Zhao, "Small-Cell Lung Cancer Detection Using a Supervised Machine Learning Algorithm", International Symposium on Computer Science and Intelligent Controls, 2017.

[3] Syed Saba Raoof, M A. Jabbar, Syed Aley Fathima, "Lung cancer prediction using Machine Learning, 2020

[4] Vaishnavi. D1, Arya. K. S2, Devi Abirami. T3, M. N. Kavitha, "Lung Cancer Detection using Machine Learning", International Journal of Engineering Research & Technology (IJERT), 2019.

[5] Ozge Gunaydin, Melike Gunay, Oznur Şengel, "Comparision of lung cancer detection algorithms", Institute of Electrical and Electronics Engineers(IEEE),2019.

[6] Naji Khosravan and Ulas Bagci," Semi-Supervised Multi-Task Learning for Lung Cancer Diagnosis", Center for Resaerch in Computer Vision (CRCV).2019.

[7] Emine CENGİL, Ahmet ÇINARA, "Deep Learning Based Approach to Lung Cancer Identification", Institute of electrical and electronics engineers 2017.

[8] Radhanath Patra, "Prediction of Lung Cancer Using Machine Learning Classifier", International Research Journal of Engineering and Technology,2020.

[9] S. Saini, A. Maithani, D. Dhiman and A. Bisht, "Analysis of Different Machine Learning Algorithms Used for Identification of Lung Cancer Disease," 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 2021, pp. 1-5, doi: 10.1109/ICRITO51393.2021.9596308.

[10] Danjuma, Kwetishe. (2015). Performance Evaluation of Machine Learning Algorithms in Post-operative Life Expectancy in the Lung Cancer Patients.

[11] S. Moreno, M. Bonfante, E. Zurek and H. S. Juan, "Study of Medical Image Processing Techniques Applied to Lung Cancer," 2019 14th Iberian Conference on Information Systems and Technologies (CISTI), 2019, pp. 1-6, doi: 10.23919/CISTI.2019.8760888.

[12] P. Lobo and S. Guruprasad, "Classification and Segmentation Techniques for Detection of Lung Cancer from CT Images," 2018 International Conference on Inventive Research in Computing Applications (ICIRCA), 2018, pp. 1014-1019, doi: 10.1109/ICIRCA.2018.8597273.

[13] N. Nawreen, U. Hany and T. Islam, "Lung Cancer Detection and Classification using CT Scan Image Processing," 2021 International Conference on Automation, Control and Mechatronics for Industry 4.0 (ACMI), 2021, pp. 1-6, doi: 10.1109/ACMI53878.2021.9528297.

eISSN 1303-5150