A Comparative Study on Machine Learning Models to Predict Lung Cancer and Types of NSCLC

Riya Adsul
Btech Computers, 4th Year Mukesh Patel School of Technology Management and Engineering (MPSTME), NMIMS University, Mumbai, India
riya.adsul57@nmims.edu.

Vedant Misra
Btech Computers, 4th Year Mukesh Patel School of Technology Management and Engineering (MPSTME), NMIMS University, Mumbai, India
vedant.misra52@nmims.edu.in

Saumya Pailwan
Btech Computers, 4th Year Mukesh Patel School of Technology Management and Engineering (MPSTME), NMIMS University Mumbai, India
saumya.pailwan42@nmims.edu.in

Abstract— **Lung cancer is one of the most common types of cancer, which is the main cause of death in humans. In order to be cured, cancer must be diagnosed at an early stage. Lung cancer, also known as lung carcinoma, is a malignant tumor that forms in the lungs and is characterized by unchecked cell proliferation. Non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC) are the two main subtypes of lung cancer. This research examines lung cancer symptoms and risk factors and uses Machine Learning algorithms to identify lung cancer patients from healthy people. These algorithms also distinguish pathological non-small cell lung carcinoma's three types. During pre-diagnosis, this classification helps choose the next step. The optimal data mining strategy is chosen by comparing its results. For the two datasets, SVM and XGBoost methods perform best.**

Keywords— Lung Cancer, NSCLC, Prediction, Data Mining, Machine Learning, Classification.

I. INTRODUCTION

Lung carcinoma, often known as lung cancer, is a type of malignant lung tumor that causes uncontrolled cell growth in lung tissues. The highest amounts of are found in North America, Europe, and East Asia. More than one-third of the highest rates are attributed to China. African and South Asian rates are significantly lower in new cases. In the United States, it is one of the leading causes of cancer-related death. It is one of the main causes of cancer-related death in men and, after breast cancer, the second greatest cause in women. Long-term tobacco use (85%), with 10 to 15% of instances occurring in people who had never smoked, is one of the contributing factors. Random gas, asbestos, smoking, and other kinds of air pollution often work together to elicit these clinical manifestations.[1] Doctors have employed a variety of methods, including screening, diagnosis, and classification, to diagnose different types of cancer at an early stage, even before symptoms occur. Furthermore, several new strategies for predicting the fate of cancer treatment in its early phases have been developed. A large amount of cancer data was collected and made available for medical research with the introduction of contemporary medical technologies. Predicting the outcome of a sickness, on the other hand, is difficult work, and it is among the most challenging functions of doctors.[2]Machine Learning has helped us by automating various tasks in various domains. It's use becomes of immense importance as it saves a lot of time and helps the doctors to focus on more problems. However, in such cases the model used must be highly accurate as it'll help in early detection of this deadly disease and save lives of people. Earlier as well lung cancer has been predicted using SVC[3].

In this paper, we aim to do a comparative study of the various machine learning algorithms used in predicting lung cancer and utilize the best model to deploy in a web application. Also, we not only detect the presence or absence of lung cancer but also the type of NSCLC if it exists. Lung cancer is classified into two types: non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC) or oat cell cancer. Lung cancer spreads and affects people differently, so treatment is also different. Mixed small cell/large cell cancer refers to cancer that has traits of both types. Smoking has been related to both NSCLC and SCLC, and it promotes tumors to grow and spread faster, resulting in big tumors that can spread throughout the body. They usually start in the bronchi in the center of the chest. NSCLC is more frequent than SCLC and develops and spreads at a slower rate. Adenocarcinoma, squamous cell carcinoma, and giant cell carcinoma are the three subtypes of NSCLC.

Adenocarcinoma: Adenocarcinomas develop from cells that normally secrete substances such as mucus. It affects women more than men and occurs at a younger age than other types of lung cancer. It is more likely to be discovered in the lungs' outermost layers before it has spread.

Squamous cell carcinoma: Squamous cell carcinomas start in squamous cells, which are flat cells that line the inside of the airways of the lungs. They are frequently associated with a smoking history and are usually found in the center of the lungs, near a major airway i.e., bronchus.
Large cell (undifferentiated) carcinoma: Any area of the lung can develop this type of cancer. It spreads and grows rapidly, making treatment more challenging. A fast- growing cancer very similar to small cell lung cancer is large cell neuroendocrine carcinoma, a subtype of large cell carcinoma.[2]

The possibility of lung cancer is identified or reduced with the help of pre-diagnosis. Symptoms and risk factors (smoking, drinking, being overweight, and insulin resistance) had a statistically significant impact in the pre-diagnosis stage. The foundation of our search strategy for potential lung cancer patients is a careful analysis of these symptoms and risk factors. It is possible to predict cancer tumors using data mining techniques. The outline of the paper is as follows: In Section II, is literature survey Section III elaborates on the data mining process and its steps. In Section IV we have given a brief insight on the working of various data mining models. Section V talks about the evaluation metrics. In Section VI we have discussed the implementation of the ML models on the datasets and discussed the results In Section VII, we have discussed the conclusion and future scope.

(978-1-6654-9291-1/22/$31.00 ©2022 IEEE)

II. LITERATURE SURVEY

In this section we have provided a brief description of the research papers based on Lung Cancer detection using various Machine Learning algorithms:

The closed structure of lung regions is depicted in the paper[4], and the disordered data is then used to generate a learned self-organizing map (SOM) using various clustering approaches. When compared to direct data clustering, the two-stage procedure, which uses SOM to generate prototypes that are then clustered in the second stage, performs well and reduces computation time.

Multilayer Perceptron, Neural Network, Decision Tree, Naive Bayes, Gradient Boost Tree, Support Vector Machine, Random Forest, and Majority Voting are all tested and compared in[5], with the latter's primary objective being to draw comparisons between them. K- factor cross validation on the UC Berkeley Lung Cancer dataset showed that Gradient Boost Tree performed better than the other classifiers tested.

Bagging is proposed for predicting lung cancer stages in[6] by layering K-Nearest Neighbors, Decision Trees, and Neural Networks using textual data. Bagging has been found to be an effective way to boost the functionality of sole models. It has been established that the accuracy of the group model is 98%. Using the Lung Cancer dataset from the University of California, Irvine, and the Data World source, [7] obtains a highest accuracy of 96.9% (Logistic regression).

CT (Computed Tomography) may aid in the detection of enlarged lymph nodes that may be cancerous and could reach other parts of the body. [8], [9] used CT images to detect the presence of cancerous nodules in the lungs.

In [10] Numerous segmentation algorithms, such as Naive Bayes and the Hidden Markov Model, were discussed. The use of different segmentation algorithms for the detection of lung cancers is thoroughly explained. Whereas [9] proposes automatic lung cancer detection using image processing of CT images, specifically Sobel and Morphological operations with high precision and cancer area detection.

SVM is employed in this study to forecast the onset of lung cancer. The main goal of this system is to give customers an early warning so they can save time and money. Positive results from the performance evaluation of the suggested approach show that oncologists can use SVM to identify lung cancer with effectiveness. If the forecast is accurate, the doctor might be able to create a more effective prescription and provide the patient an earlier diagnosis.[11]

The numerous models used by researchers were laid out along with their accuracy levels, and they have a few limits and drawbacks that were noted. A thorough review of the literature revealed that certain classifiers have poor accuracy while others have higher accuracy but have not yet approached 100%. Tried using Backpropagation network, image segmentation.[12]

In the majority, CT images from scans are used to find cancer. Additionally, marker-controlled compared to other methods,watershed segmentation yields more precise results. Segmentation strategies as a bonus, the outcomes of the methods based on deep learning produced more accuracy than the techniques that have been used with traditional machine learning techniques. [13]

Another approach for early detection was done using U-net architecture which gave much higher accuracies for CT scan images. The work proposed a 3D multipath similar to VGG network which is evaluated on extractions of images from database. The lung nodules were classified, and accuracy detected is around 95.6% which is much higher than existing other methods.[14]

Various commonly used deep learning methods for Lung cancer include Convolution Neural Network. The work comprises of comparing this with other methods along with their advantages and disadvantages. Basically, comparing the conventional CAD systems with the methods established on Deep Learning technique.[15]

The dataset available in open source is also used and trained using a variety of techniques, including Support Vector Machine (SVM), K-Nearest Neighbor, Decision Tree, Logistic Regression, Naive Bayes, and Random Forest, and it is demonstrated that these approaches are more accurate. The Random Forest method has generated improved performance with 88.5% accuracy.[16]

None of the papers implement the detection of lung cancer and the identification of the type of lung cancer consecutively using the symptom and risk factor data of patients. Furthermore a 98.14% test accuracy is achieved for the classification of Lung Cancer in this research. Using the symptoms along with visual data can result in a more accurate detection of the disease.

A comparative analysis of some of the works is shown below:

TABLE I. COMPARATIVE ANALYSIS

Ref	Preprocessing	Algorithms	Dataset	Results
[5]	Missing values, outlier detection etc. 10-fold cross-validation	Auto MLP, Naïve bayes, SVM, Decision Tree, Majority voting, Neural Network	UCI online ML repository	SVM and Decision Tree outperforms rest classifiers
[6]	RGB to greyscale, binarization, segmentation, feature extraction	Random Forest, SVM	CT scan medical images	Random Forest gives 66% efficiency and SVM gives 94.5%
[8]	PCA, Eigen Vectors	Naïve Bayes, ANN, SVM, KNN	JSRT	SVM – 97% ANN – 96%
[9]	Converting to grey scale images, Image filtering, enhancement, Morphological operations, and Feature Extraction	Multilayer perceptron, SM and Radial basis function neural network. Proposed Stacked sparse auto encoder, SSAE	CT scan medical images	MSE (0.177), RMSE (0.172) and MAE (0.164)
[2]	EDA and feature selection using SelectKBest method, tree-based method	XGB, Logistic Regression, SVM, Decision Tree, KNN, Gaussian NB	Lanzhou University in Second Hospital, China	XGBoost highest accuracy 95%

III. DATAMINING APPROACH

Data patterns are extracted from data using data mining techniques. The data mining technique employed determines the patterns that can be found. Data mining activities are typically divided into two categories: descriptive and predictive. Descriptive tasks describe the general qualities of existing data, while predictive tasks try to generate predictions based on the data that is currently available. Some of the steps involved in data mining areas follows:

Data exploration: If the data quality is not sufficient for a dependable model, recommendations for improved information collection and cache might be made. Knowledge from many sources must be merged for analysis to ensure uniform treatment.

Data preparation: This step involves cleaning and transforming the data to ensure that all known valid values are consistent, and that missing and invalid values are handled for a more thorough analysis.

Modelling: Based on the data and desired results, a data mining method or combination of algorithms is selected for analysis. These algorithms combine cutting-edge methods with more established ones, such as statistics, neighborhoods, and clustering. The algorithm is chosen based on the specific goal to be achieved as well as the quality of the data to be analyzed.

Evaluation: Based on the findings of the data mining algorithms, a study is carried out to identify key findings and develop a list of suggestions for consideration.

IV. METHODOLOGY

In this section we will discuss the working of a few important machine learning models that we have used for the classification of lung cancer further in the paper.

A. Naïve Bayesian

Naive Bayesian is a method of classification that makes use of Bayes' Theorem and the independence condition for predictors. A Naive Bayes classifier, in its most basic form, treats the presence or absence of a feature inside a class as independent of the presence or absence of any other characteristic. It determines the likelihood that a given item or data point corresponds to a particular class by computing affiliation odds for each class. Most people believe that the most probable category is the one that has the highest proportion.

B. Logistic Regression

Based on the values of independent variables, which can be categorical or numerical, logistic regression attempts to predict the likelihood of an event occurring. It is a statistical method used to analyze data sets that contain one or more independent variables that influence the outcome. The goal of this method is to find the best-fitting model that describes the relationship between a set of independent (predictive) variables and the dependent variable. Our problem is both a single-class classification and a multi-class classification problem. The fundamental equation of generalized linear model is:

$$g(E(y)) = \alpha + \beta x1 + \gamma x2 \qquad (1)$$

Here, g () is the link function, E(y) is the expectation of target variable and $\alpha + \beta x1 + \gamma x2$ is the linear predictor (α, β, γ to be predicted). The role of link function is to 'link' the expectation of y to linear predictor.

C. Decision Tree

An efficient machine learning method for predictive analysis is the decision tree. By creating decision rules, it enables us to analyze data quickly and accurately. We can forecast the target variable after segmenting the dataset according to the different decision rules. A greedy top-down recursive divide-and-conquer strategy is the foundation of this algorithm. The leaf node designates the class, and the internal nodes are named after attributes, with the outgoing edges representing the outcomes of test conditions on that attribute. Based on the value of the attribute, nodes with attributes split the training dataset into two or more subgroups. Attribute selection measure selects the attributes that best divide the training dataset at each stage. Following this, the decision tree model performs cyclical operations on the residual attributes and data subsets on the remnant outgoing edges. If all tuples associated with the same class are encountered, the leaf node is tagged with that class; anything other than that, the leaf node is tagged with the majority class of tuples (if the attribute list is empty) or the hugely influenced class of the node's primary training tuples (if no tuples are found).

D. Random Forest

Random forests, alternatively referred to as random decision forests, are an ensemble learning strategy for classification that involves training numerous decision trees synchronously and then generate the class that corresponds to the majority of the class features extracted by individual trees. The problem of overfitting in decision trees is addressed by random decision forests. At each fork, discrete decision trees are compiled using scrambled subsets of tuples and attributes from the pertaining training datasets. Leveraging bagging and a stratified random of characteristics, a Random Forest may be devised. To teach tree learners, the random forest training method employs the bagging approach. The following bagging technique is applied to a single tree, but random forest takes a different approach, selecting a random subset of features at each split in the learning process, which is often referred to as "feature bagging." Where one of the traits is a reliable predictor of class, this is performed to minimize the trees from correlating with one another.

E. Support Vector Machine

The classification methodology known as support vector machine learning (SVM) is grounded on supervised learning but is not subject to random variation. Specifically, it seeks to provide the optimal line, or decision boundary, for identifying n-dimensional space, so that fresh data points may be efficiently categorized in the future. The most useful boundary for making a call is a hyperplane. For creating the hyperplane, SVM pursues foremost severe points/vectors. The vectors of assistance are the selections from the margins. In order to train the SVM, we need information from both classes.

F. Linear Discriminant Analysis

Though effective for binary classification, the linear classification model known as logistic regression has limitations when used to problems involving numerous classes that are clearly differentiated from one another. Although LDA performs admirably with these. In Machine Learning and other pattern classification applications, Linear Discriminant Analysis is a common method for reducing dimensionality used as a pre-processing step. It is a label- based, supervised classification method. Reducing the number of dimensions is the fundamental objective of dimensionality reduction methods, which accomplish this by mapping features from a higher-dimensional space to a lower- dimensional space, hence eliminating superfluous or correlated features.

G. Voting Classifier

A Voting Classifier is a type of machine learning method that adapts from a broad range of models and then leverages the model with the best likelihood to predict a result class. It simply aggregates the predictions of all classifiers pumped into Voting Classifier and generates the final prediction based on the class that garners the largest majority. By incorporating the predictions of numerous models into a single, unified model, we eliminate the need to forge and investigate individual models for each target result. Voting Classifier supports two types of voting: hard voting and soft voting.

H. Gradient Boost

To create a prediction model, gradient boost uses a collection of relatively weak prediction models for regression and classification. This procedure builds a model incrementally and then generalizes it such that any differentiable loss function may be optimized. The goal of the iterative procedure known as "gradient boost" is to combine several weak classifiers into a single robust one.

I. XGBoost

The underpinning for this approach is gradient boosting. It is a strategy for assembling and utilizing Decision Trees. Boosting is a collaborative method in which each tree alters the evaluation standards based on the information it receives from the tree that came before it. This enhances the overall efficiency of the model by aggregating and transforming weak classifiers into strong ones. Gradient boosting is followed by XGBoost, which uses optimization techniques to provide better results in a short period of time.

V. EVALUATION METRICS

We have considered two measures to evaluate the effectiveness of the various classification techniques under consideration. Accuracy is the first metric considered in the data (train and test). In general, accuracy is an indicator of how frequently the classifier correctly identifies the class labels. Accuracy is gauged as follows:

$$\textbf{accuracy} = (TP + TN) / (P + N) \quad (2)$$

TP stands for 'true positive,' whereas TN indicates a 'true negative. The true positive rate, P, is the entire number of positive instances in the data, whereas the true negative rate, N, is the whole number of negative cases. Referring to the lung cancer data, it refers to yes, no and 1,2,3 target class. If a test's expected result is YES however the expected result is NO, for instance, the test's outcome won't factor into TP and TN. A false positive, however, will be recorded.

The F1 score is the second statistic we've employed. A test accuracy can be quantified with an F1 score (also known as an F-score or F-measure). To arrive at a final grade, it takes into account both the accuracy and the overall memory of the system. The proportion of accurate positive results, p, is the same as the proportion of all positive results, n, and the proportion of expected positive results, r, is the same as the proportion of all expected positive results, n.. Precision is given by:

$$\textbf{Precision} = TP/(TP+FP) \quad (3)$$

Where TP is true positive, and FP is false positives. Recall is given by:

$$\textbf{Recall} = TP/(TP+FN) \quad (4)$$

Where TP is true positive, and FN is false negatives. Precision, then, is the proportion of times our 2 declarations were correct relative to the total number of times the algorithm made a 2 declaration. Conversely, recall is the proportion of occasions where we accurately announced 2 out of all the cases when the true state of the world is 2. The F1 score may be understood as a combination of the important levels assigned to accuracy and recall, with a perfect score being 1 and a failing score being 0.

VI. IMPLEMENTATION AND RESULTS

The section compares the performance of certain fundamental classifiers and talk about the Lung Cancer data - set, experiments, and assessment system. The results are shown in the diagram below.

Our first goal is to differentiate between patients with lung cancer and those who are healthy based on a variety of factors and symptoms. Our next aim is to identify which type of pathological lung cancer (NSCLC) is the person suffering from namely (Lung Adenocarcinoma, Squamous Cell Carcinoma, Large Cell Lung Cancer). The implementation is divided into the following 6 steps:

Step-1: Dataset description. For differentiating between healthy person and lung cancer patients Dataset retrieval is done through an online archive.(data. world)

Dataset 1: UCI Machine Learning Repository has been used for acquiring the dataset.

Attribute Information: Total no. of attributes = 16 (1 class attribute,15 predictive) and No. of instances = 284. Attributes, which stand in for the symptoms themselves, are used to great efficacy in lung cancer categorization. This dataset's attributes are listed in the table below:

TABLE II. DATA.WORLD DATASET

Attribute	Description
Gender	M(Male), F(Female)
Age	Age of patient
Smoking	YES=2, NO=1
Yellow fingers	YES=2, NO=1
Anxiety	YES=2, NO=1
Wheezing	YES=2, NO=1
Alcohol	YES=2, NO=1
Coughing	YES=2, NO=1
Shortness of breath	YES=2, NO=1
Swallowing difficulty	YES=2, NO=1
Chest pain	YES=2, NO=1
Lung Cancer	YES=2, NO=1

For identifying the type of pathological lung cancer (NSCLC) the patient is suffering from.

Dataset 2: Dataset retrieval is done through an electronic archive. (uci). The data described 3 types of pathological lung cancers. The Authors give no information on the individual variables nor on where the data was originally used. Number of Instances: 170, Number of Attributes: 57 (1 class attribute, 56 predictive).

Attribute Information: attribute 1 is the class label. And all predictive attributes are nominal, taking on integer values 0-3.

Step-2: **Preprocessing of Data**

Dataset 1: Application of pre-processing for data cleaning and modification – Dataset present is imbalances with count of values for class 1 being 270 and for class 0 being 39. To balance out the dataset Scikit learn's resampling function is utilized. Minority upsampling is performed having random state parameter set to 4 and number of samples set to 269.

Dataset 2: Application of pre-processing for data cleaning- the dataset is perfectly balanced, so resampling is required

Step-3: Calculating Correlation Matrix

Dataset 1: Calculating the correlation and matrix and plotting the heatmap – The attributes having the highest correlation with lung cancer are "allergy", "alcohol consumption" and "swallowing difficulty".

Dataset 2: Calculating the correlation and matrix and plotting the heatmap for the multiclass data.

Step-4: Train Test Split

The given datasets are divided into training and test data with test size set to 0.2. Standardization of data is performed using StandardScaler to render the data scale free to avoid biased outcomes.

Step-5: Implementation of ML Models

Machine learning models with appropriate hyperparameters are trained using training data. After developing the models, they are put to the test with real-time (test) data and their performance is measured against a set of predetermined criteria. The models used in the experiment are:

Linear Discriminant Analysis, Logistic Regression (with solver "lbfgs" and max iteration 200, SVC with C value set to 1 gamma 1 and kernel "rbf", Random Forest with a max depth of 5, Decision Tree with criterion entropy, Gaussian Naïve Bayes, KNN with 7 neighbors', Gradient Boosting Classifier with learning rate set to 0.01 and random state 1,AdaBoost, XGBoost with learning rate 0.01 and random state 1, MLP with 800 hidden layers and random state 50 and Voting Classifier using Decision Tree and SVC).

Step-6: Comparing the Results & Identifying the Best Classifier

The prime classifier for identifying lung cancer is determined by comparing the performances of all available classifiers, both standalone and in ensembles

TABLE III. COMPARISON OF RESULTS (DATASET 1)

Sr No	Algorithm	Testing Accuracy	Training Accuracy
0	Linear Discriminant Analysis	0.87963	0.916
1	Logistic Regression	0.87037	0.914
2	AdaBoost	0.925926	0.965
3	SVC	0.7981481	0.997
4	Random Forest	0.916667	0.944
5	Decision Tree	0.953704	0.997
6	Gaussian NB	0.888889	0.861
7	K Neighbors	0.851852	0.883
8	Gradient Boosting	0.861111	0.907
9	XGB Classifier	0.861111	0.897
10	Voting Classifier	0.953704	0.997
11	MLP Classifier	0.916667	0.909

Table III. shows that the performance of Support Vector Machine Classifier exceeds the performance of Decision Tree and all others (98.14% test accuracy, precision 96.3%, recall 100% and F1 score 98.1%).

The scikit-learn function k- fold cross valuation score (with a 4-fold validation, decision function "ovr", chache size 200, gamma 1 and kernel "rbf") is calculated for the SVM classifier to evaluate (train and test) it over several dataset folds. Instead of only a train/test split, cross validation will help us understand it's performance across the entire dataset. The accuracy obtained after cross validation for SVC is 0.986. So, we can conclude that SVM has the highest accuracy rate among all other classification algorithms for this dataset1.

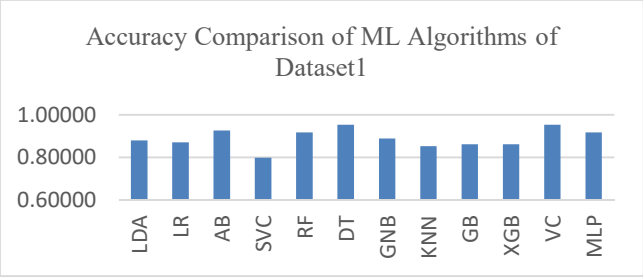

Fig. 1. Graph showing Comparison of classifiers (Dataset1)

TABLE IV. COMPARISON OF RESULTS (DATASET 2)

Sr No	Algorithm	Testing Accuracy	Training Accuracy
0	Linear Discriminant Analysis	0.714286	0.875
1	Logistic Regression	0.571429	1
2	AdaBoost	0.428571	0.708
3	SVC	0.571429	0.9583
4	Random Forest	0.571429	1
5	Decision Tree	0.571429	1
6	Gaussian NB	0.285714	0.9166
7	K Neighbors	0.714286	0.666
8	Gradient Boosting	0.714286	1
9	XGB Classifier	0.857143	0.958
10	Voting Classifier	0.714286	1
11	MLP Classifier	0.714286	1

The data in the table strongly implies that the performance of XGBoost Classifier (85.71% test accuracy) which used (random state=1, learning rate=0.01) as hyperparameters exceeds the performance of KNN, MLP and all others. We may thus infer that XGBoost outperforms all other classification methods on this dataset with respect to accuracy.

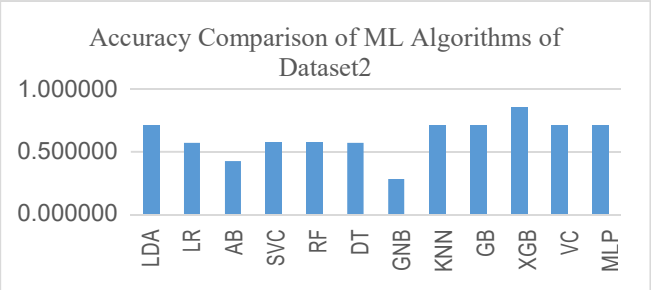

Fig. 2. Graph showing Comparison of classifiers (Dataset2)

VII. DISCUSSION & CONCLUSION

Previously, a multitude of tests were required to detect if or not a patient seemed to have lung cancer. Nevertheless, this was a cumbersome process. In order to diagnose lung cancer, an individual may be constantly exposed to needless examinations or procedures. To decrease process time and fruitless screenings, a preliminary test informing the patient and physician of the likelihood of lung cancer should be conducted. From the accuracy of the data, using Machine Learning techniques, and with access to additional data, we may draw the conclusion that cancer can be anticipated, it is possible to detect tumors and forms of cancer without using CT scans, using only textual data. There are benefits and drawbacks to using any given algorithm. Naive Bayes has good performance and is faster if there is independence between qualities, but it is not always the case. Although they are the most time-consuming, multilayer neural networks excel at processing numerical input and can deal with chaotic information. When expressed as a decision tree, the categorization model is both faster and easier to comprehend. Random forest enhances decision tree accuracy by addressing the issue of overfitting to training datasets that plagues traditional decision trees. Better accuracy is achieved, and linearly inseparable data may be processed using SVM since the difficulty of the learnt classifier is determined by the number of support vectors as opposed to the dimensionality of the data. When given noisy or irrelevant qualities, k-nearest neighbor learns slowly and inaccurately; however, accuracy can be increased by feature extraction.

In order to conduct this analysis, many machine learning methods were used. Each classifier's accuracy is analyzed. Classifiers are compared quantitatively with regards to their ability to predict. Various classifiers generate different outcomes in the performance chart when applied to the lung cancer dataset. Both the support vector machine method, which is used to classify lung cancer, and the XGBoost algorithm, which is used to identify the specific kind of pathological lung cancer, give the best results in terms of proper classification and other metrics. The SVM algorithm classified the observation using a high dimension, resulting in the highest results. This method permits for more precise lung cancer diagnosis. XGBoost, in contrast, is predicated on gradient boosted decision trees and may be used to solve classification issues when the dataset is big (well over 1000 rows) and contains missing values along with categorical and numeric variables. XGBoost is capable of handling anomalies, missing values, and unscaled data, and may be used to enhance model effectiveness and overall execution speed. In the end, the accuracy rate may be raised by adding more pre-processing and hyperparameter adjustment. There is room for growth and development in the method for predicting lung cancer. Methods from the field of data mining, such as Time Series analysis, clustering, and the application of association rules, are also applicable. It is possible to employ continuous data alongside to categorical data. Medical databases include vast volumes of unstructured data that may be mined using Text Mining. Adding to the complexity, integrating data mining with text mining would be a significant challenge. We can also use symptoms and CT scan lung images to train a deep learning model to provide better results in all scenarios.

REFERENCES

[1] E. Y. V. Chandra, K. R. Teja, M. H. C. S. Prasad, and B. M. Ismail, "Lung cancer prediction using data mining techniques," *International Journal of Recent Technology and Engineering (IJRTE)*, vol. 8, no. 4, 2019.

[2] A. Chauhan and others, "Detection of lung cancer using machine learning techniques based on routine blood indices," in *2020 IEEE international conference for innovation in technology (INOCON)*, 2020, pp. 1–6.

[3] S. S. Raoof, M. A. Jabbar, and S. A. Fathima, "Lung Cancer prediction using machine learning: A comprehensive approach," in *2020 2nd International conference on innovative mechanisms for industry applications (ICIMIA)*, 2020, pp. 108–115.

[4] H. Bharathi and T. S. Arulananth, "A review of lung cancer prediction system using data mining techniques and self organizing map (SOM)," *International Journal of Applied Engineering Research*, vol. 12, no. 10, pp. 2190–2195, 2017.

[5] M. I. Faisal, S. Bashir, Z. S. Khan, and F. H. Khan, "An evaluation of machine learning classifiers and ensembles for early stage prediction of lung cancer," in *2018 3rd international conference on emerging trends in engineering, sciences and technology (ICEEST)*, 2018, pp. 1–4.

[6] K. Roy *et al.*, "A Comparative study of Lung Cancer detection using supervised neural network," in *2019 International Conference on Opto-Electronics and Applied Optics (Optronix)*, 2019, pp. 1–5.

[7] P. R. Radhika, R. A. S. Nair, and G. Veena, "A comparative study of lung cancer detection using machine learning algorithms," in *2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, 2019, pp. 1–4.

[8] Ö. Günaydin, M. Günay, and Ö. Şengel, "Comparison of lung cancer detection algorithms," in *2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)*, 2019, pp. 1–4.

[9] S. Baskar, P. M. Shakeel, K. P. Sridhar, and R. Kanimozhi, "Classification system for lung cancer nodule using machine learning technique and CT images," in *2019 International Conference on Communication and Electronics Systems (ICCES)*, 2019, pp. 1957–1962.

[10] A. R. Kaur, "A study of detection of lung cancer using data mining classification techniques," 2001.

[11] C. Anil Kumar *et al.*, "Lung Cancer Prediction from Text Datasets Using Machine Learning," *Biomed Res Int*, vol. 2022, 2022.

[12] E. S. N. Joshua, M. Chakkravarthy, and D. Bhattacharyya, "An Extensive Review on Lung Cancer Detection Using Machine Learning Techniques: A Systematic Study.," *Rev. d'Intelligence Artif.*, vol. 34, no. 3, pp. 351–359, 2020.

[13] D. M. Abdullah, N. S. Ahmed, and others, "A review of most recent lung cancer detection techniques using machine learning," *International Journal of Science and Business*, vol. 5, no. 3, pp. 159–173, 2021.

[14] R. Tekade and K. Rajeswari, "Lung cancer detection and classification using deep learning," in *2018 fourth international conference on computing communication control and automation (ICCUBEA)*, 2018, pp. 1–5.

[15] S. Das and S. Majumder, "Lung cancer detection using deep learning network: A comparative analysis," in *2020 Fifth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)*, 2020, pp. 30–35.

[16] S. Bharathy, R. Pavithra, and others, "Lung Cancer Detection using Machine Learning," in *2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, 2022, pp. 539–543.