https://doi.org/10.33472/AFJBS.6.11.2024.1679-1695

African Journal of Biological Sciences

Journal homepage: http://www.afjbs.com

Research Paper Open Access

PERFORMANCE EVALUATION OF BOOSTING ALGORITHMS BASED MACHINE LEARNING MODELS FOR PREDICTING LUNG CANCER

[1]Venkat P. Patil, [2]Pravin R. Kshirsagar, [3]Bhuvan Unhelkar, [4]Prasun Chakrabarti

[1]Smt. Indira Gandhi College of Engineering, Navi Mumbai, India,
[2]J D College of Engineering & Management, Nagpur, Maharashtra, India,
[3]Muma College of Business, University of South Florida, Sarasota campus, USA,
[4]Sir Padampat Singhania University, Udaipur Rajasthan India.

Corrousponadance author : Venkat P. Patil , Email : Venkat.patil@sigce.edu.in

Article Info

Volume 6, Issue 11, July 2024

Received: 21 May 2024

Accepted: 27 June 2024

Published: 12 July 2024

doi: 10.33472/AFJBS.6.11.2024.1579-1695

ABSTRACT:

Among those most fatal conditions that necessitate early detection nowadays is lung cancer. Because it can lessen the likelihood of human mistake while evaluating medical images, artificial intelligence has become an indispensable tool in the medical industry and, more specifically, in the analysis of medical images and the diagnosis of diseases. The rapid advancement of machine learning (ML) algorithms for prediction has revolutionized various industries, including medical treatment, facilitating the effortless early detection of lung cancer. Machine learning algorithms have the ability to forecast or diagnose a wide range of serious ailments, including cancer, lung cancer, heart disease, etc., many of which can be dangerous. In this paper, we examine and contrast numerous Machine learning methods that use Boosting algorithms for predicting the onset of diabetes in its early stages. The major objective of this research work is to establish the most efficient classifier for lung cancer detection by organizing and carrying out the procedure using many Boosting based machine learning (ML) techniques. In this study, we examine a broad variety of disease-related traits in an effort to provide a more complete picture of lung cancer and its prognosis. In this research, we employ numerous Boosting algorithm-based Machine Learning classifications strategies to the traditional Lung Cancer Dataset. These techniques include Gradient Boost (GB), XGBOOST (XGB), ADABOOST, CATBOOST (GB), and LightGBM (LGBM). When it comes to accuracy, the models employed here are all over the map. This study demonstrates a method that may reliably forecast the occurrence of lung cancer. This study's findings suggest that the GB Model, a machine learning classifier belonging to the class of Boosting algorithm-based models, is the most effective in predicting the occurrence of carcinoma of the lung.

Keywords—Lung Cancer prediction; Gradient Boost, XGBOOST (XGB), ADABOOST, CATBOOST (GB), and LightGBM.

1. INTRODUCTION.

Early categorization and prognosis is crucial in this healthcare sector for complicated medical assessment and treatment prediction. Machine learning and deep learning have advanced to the point that they can even forecast the complexity and development stages of cancerous cells. Cancer is the most dangerous illness in medicine since early detection is key

to simplifying prognosis and therapy. Focusing on some of the important aspects of lung cancer diagnosis and therapy categorization using DL methods is the main focus of the current research. Cancers that start in the lungs are known as lung cancers. Inhaling causes your porous lungs in your chest to emit carbon dioxide and inhaling causes them to take in oxygen. More people die from lung cancer every year than from any other cancer in the body. While cigarette smokers are at a higher risk of developing lung cancer, anybody may get the disease. Both the duration and quantity of cigarette smoking are associated with an elevated risk of lung cancer. Even if you've smoked for a long time, you may greatly lower your risk of developing lung tumors by quitting. It is of the utmost importance to discover this malignancy quickly.
Boosting is the process of applying the underlying learning algorithm to modified input data repeatedly [1]. Boosting algorithms make use of input data for training a weak learner, compute the learner's predictions, further select misclassified training samples, and then train the subsequent weak learner with an updated training dataset that incorporates the instances that were incorrectly classified during the previous training cycle [2].

Among the main aims of this study was to determine if or whether
a. There are any publicly available datasets used for lung carcinoma investigation.
b. Carried out a comprehensive investigation performance Evaluation of different boosting-based Machine learning models like Gradient Boost (GB), XGBOOST (XGB), ADABOOST, CATBOOST (GB), and LightGBM (LGBM) methods.
c. Using performance indicators to analyze the success of prompt lung cancer diagnosis.
Part II of the investigation study reviews the corresponding publications; it follows the introduction and overview. Section III offers a concise synopsis of the methods and processes that comprised our research. Experimental findings are presented in Section V, while the suggested technique and metrics for assessing performance are detailed in Section IV. The last part, VI, emphasizes the findings and the conclusion that follows.

2. RELATED WORKS

Boosting can help weak classifiers improve. Meta-algorithm ensembles reduce bias and variation. Strong learners form the basis of boosting ensemble algorithms, whereas weak learners are classifiers that outperform random guessing [3]. In response to Kearns and Valiant's [4] concern about whether a group of Learning from less capable learners might result in a more capable learner. Schapire [5] developed the boosting method in 1990. AdaBoost [6] and XGBoost [7] were inspired by Schapire's [5] work on machine learning and statistics. What follows is a discussion of some research projects that have used machine learning techniques to either detect or predict the occurrence of lung carcinoma.
A DL framework for computational detection of lung tumours in chest CT scans was created and verified by himazaki et al. [8]. There were 629 images in the preliminary population with 652 nodules/masses, and 151 images in the validation dataset with 159 nodules/masses. The findings from the model in the independent test dataset were as follows: sensitivity = 0.73, mFPI = 0.13. On the other hand, as opposed to nonoverlapped sites, the model's response rate was less for lung tumours that overlapping with blind spots.
An approach for the detection of lung_cancer using DL residuals on “CT-Scan” pictures was built by Bhatia et al. [9]. In order acquire characteristics and recognize areas that may be susceptible to cancer, the investigators used ResNet and U-Net algorithms. The cancer forecasting process made use of a number of machine learning algorithms, such as XG boost, RF, and individualized forecasts. On the LIDC-IDRI data set, the method attained a success rate of 84%.

To successfully detect lung tumours, Talukder et al. [10] suggested a hybrid ensemble feature extraction approach. The LC25000 dataset, which contains information on the lungs, was used to test the framework. Those findings demonstrated that the mixed approach was able to identify lung cancer with an astounding 99.05% effectiveness rate. These findings prove that the suggested method is useful for making precise diagnoses of lung cancer. The research also shown that the suggested mixed approach far surpassed the current models, suggesting that it may be useful in clinical situations. In order to improve the efficiency of lung cancer detection, this proves that combination models of TL approaches work.
In 2021, Wang et al. suggested a method determined by RNNs for the detection and treatment of lung carcinoma by analysing longitudinal information from EMRs. Forecasting the likelihood of getting lung cancer up to a year before it happened was a strong suit of the RNN model [11]. For the purpose of identifying lung nodules and early-stage lung cancer on CT images, Liu et al. (2020) suggested a CNN-SVM hybrid model. Showing the promise of merging many NN classifications for enhanced performance, the model attained great accuracy in detecting lung nodules and diagnosing lung tumours [12]. Making use of fuzzy clustering-based decision trees technique, Mouttham et al. (2020) suggested a framework for lung cancer detection. The system outperformed competing deep learning approaches and reached a high level of accuracy when applied to CT scans for the diagnosis of lung carcinoma [13].

Marjolein A. Heuvelmans et al. [14] created the CNN for Lung Cancer Prediction to distinguish healthy tumours from malignant ones yet maintain the sensitivity level high. The machine learning model was fed an unbiased dataset of ambiguous nodules in a European multi-centre trial. It has been developed employing screened data from the US before. A total of 2106 nodules, including 205 lung carcinomas, were used to verify the LCP-CNN. This validation was conducted as part of the Early Lung Cancer Diagnosis study.
As regards to Many ML and DL, the study in question made use of classification techniques to boost effectiveness. For the most part, studies used like performance metrics like “accuracy, recall, F-score, precision, ROC-score, and execution time” as a means of contrasting and determining the best approach.

3. METHODOLOGY

The primary goal of this investigation is for finding out the best classifiers for lung cancer prognosis by organizing, implementing, and analyzing the outcomes of several Machine Learning techniques. Following this, we will go over the steps quickly. Figure 1 provides an overview of the proposed method for predicting the occurrence of lung tumours.

3.1. Dataset Description

In this research work, we rely on Classic Publicly available Lung Cancer Data Set be accessed from Kaggle website (Dataset, Lung Cancer Data Set). The dataset in question has 309 items and 16 features, one of which is a lung cancer-related trait. Out of the 309 reports, 270 have been "tested positive," meaning the patient has lung cancer, and 89 have been "tested negative," meaning the patient cannot have lung tumors.

3.2. Data Pre-processing

It is crucial to pre-process data. This approach guarantees reliable outcomes and accurate dataset predictions for ML algorithms (Soni et al., 2020) [3]. Although a few non-essential attributes in the Indian Lung Cancer dataset contain zero values, there are not any missing (NAN) values overall. Despite a few non-essential attributes in the Indian Lung Cancer dataset contain zero values, there are no missing value (NAN) values overall. We calculate the required average and median values for each column with zero values for Lung Cancer and no-Lung Cancer patients. Patients with and without diabetes both enter "0". For verification and training,

we used 75% of the standardized Lung Cancer Data Set, whereas for evaluation, we used 25%. The Python programming language is the language that the model is programmed in.

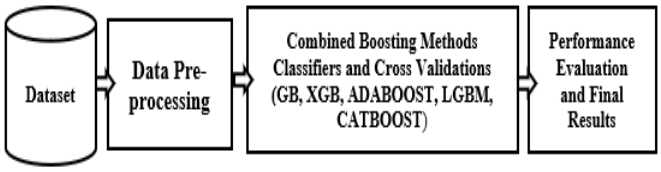

Figure 1. Proposed Model

3.3. A General Introduction to Boosting Methods: GB, CatBoost, GB, AdaBoost, Light Boost:

The goal of learning in groups or ensembles approach is to train the model employing a wide variety of computational methods. An instance of collaborative learning is the Bagging approach, which involves running different models on separate subsamples of the same dataset all at once. The boosting strategy is frequently used in practice and aims both the approach and the model will be trained upon, contrary with parallel building. We use a straightforward method to train the model, and then we reorganize it based on the results so that it can learn better. The next method receives the modified model and uses the simplified learning to its advantage. Several different boosting methods, each with its own spin on the sequencing method, are detailed in this paper.

We classify our collection of data using Machine Learning algorithms once it is prepared. We use characteristics such as smoking, anxiety, yellow fingers, chronic disease, fatigue, and XGBoost, GB, and LGBM algorithms in this study. Characteristics identified from the dataset employing Exploratory Data Analysis include: Allergy, Wheezing, Alcohol_Consumption, Coughing, Shortness_of_Breath, Swallowing_Difficulty, Chest_Pain, LungCancer, and two more characteristics.

3.3.1 XGBoost: XGBoost (XGB)

A trustworthy distributed machine learning environment for scaling tree-boosting techniques, the infrastructure provides an easy-to-understand and fast implementation of the Gradient Boosted Trees method. For quick parallel tree generation, the classification algorithm is fault-tolerant and well configured for a distributed configuration. Data from a single node, which might number in the billions, is mixed with scale-beyond distributed software samples. [15].

$$y = \sum_{i=1}^{n} (w_i \cdot x_i) \quad (1)$$

3.3.2 AdaBoost

Decision stump-based boosting algorithms like Adaboost [16] are very common. Adaboost does not do this process in a robotic fashion. In order to arrive at the best possible estimates that multiple methods change their weights in a sequential fashion. Error is calculated by every approach. The weights are updated using a second technique. The second algorithm classifies the model, changes weights like the first model, and sends it to the third algorithm. This procedure is performed to the end either the total number of estimators is reached, or the error is equal to zero. By transferring updated weights to the next stage, the approach improves categorization. A complex sequential mechanism in action: Consider blue and red labels. Deficient classifier 1 mistakenly assigns a blue label to one piece of red data. The subsequent model takes these erroneous assumptions into account and adjusts the weights of the correct ones accordingly. Due to its ability to misclassify a growing amount of bias in the samples and correct them with diminishing bias, the new model learns quicker than the old technique. On to the next steps, just repeat the process. Weak categories are necessary for powerful classifications. By importing AdaBoostRegressor, regression may be accomplished.

3.3.3 Gradient Boost:

The decision stump was a recent addition to Adaboost's improved weighting mechanism, which consists of one node separated into two leaves. The sequential approach known as gradient boost [17] also optimizes the loss by producing 8 to 32 branches, which causes trees to grow larger. To get the tax loss, use the residual from the linear model. The residual error is equal to the distinction between (both of) the measured value of y and the predicted value of y, and the sum of the squares for each and every data point represents the loss. What does the square represent? Forecasting errors are crucial since the target value is the discrepancy that exists between projected and actual values. Squaring a negative number results in a little loss, regardless of whether the value is zero or not, hence negative numbers are squared. To summarize, the subsequent algorithm receives a collection of residual values, which are reduced before being passed on to the next approach.

3.3.4 ***LightBoost (LGBM)***

The "Light Gradient Boosting Machine" (abbreviated as LGBM) is a decision tree-based Gradient Boosting method that Microsoft announced in 2017. Unlike earlier methods, it can precisely find and disable opposing soldiers by dividing the tree according to the leaves. LGBM is an effective method for increasing speed and accuracy while decreasing the likelihood of error. When utilizing the customized approach to split data into categories, it is necessary to substitute a numerical value (like an index) for the column's text name.

3.3.5 ***CatBoost***

In 2017, Yandex developed CatBoost. One-Hot-Encoding is the root of categorical boosting as it numerically converts all categorical attributes [19]. You might also put the index value next to the column name. Missed numbers may be accommodated. As compared to XGBoost, it functions better. In contrast to other boosting methods, Catboost employs symmetric networks where the number of nodes at each level is equal. Both XGBoost and LGBM train the model to a residual goal value by computing the amount of leftover error for every single data item through repeated training, it reduces the residual error until it achieves the goal.

Because this method is applied to every single data point, excessive fitting and poor generalization are both possible outcomes. Catboost will create residual for each data point by applying the model it has trained with to a number of previous data points. Individual the leftover information is produced by every data point. With each evaluation of this data, the generic model is trained again. Since several models will be used, this calculation will be time-consuming and expensive. Boosting that is organized takes less time. The chronological sequence of the data elements is used to begin ordered boosting rather than the individual data elements' residuals (n+1). Calculate n+2 by applying n+1.

4. PROPOSED METHODOLOGY AND PERFORMANCE METRICS

We began by cleaning up the Lung Cancer dataset so it could be used in our analysis. The dataset was pre-processed using tenfold cross-validation, which then divided it into a train set and a test set. Then, the training set is mostly used to detect early stages of lung cancer mellitus using the suggested ways. The last step is to evaluate performance on the test set by making use of evaluative mechanisms. We will touch on these eras quickly in the following section.

4.1 Dataset and Attributes

The purpose of this study was to investigate the efficacy of boosting based Machine Learning techniques in the early stages of lung cancer identification by using the Lung Cancer dataset. Indications, both positive and negative, are used in order to determine the patient's likelihood of getting lung cancer. These include sixteen features, ranging 89 of which are favourable and 270 of which are unfavourable.

4.2 ***Pre-Processing:***

As part of the data pre-processing that enabled us for achievement of our research goal, we addressed values that were missing in the data that had already been processed. As an example, minimum value assumptions about attributes are not appropriate for use in lung cancer prediction utilizing deep learning and machine learning. For example, we assign a 1 to "yes" and a 0 to "no." This allows us to determine nominal characteristics like "male" and "female" in the Gender category, "yes=1" and "no=0" in the other attributes category, and "positive" and "negative" in the class groupings, such as Lung Cancer, by using these numerical representations.

4.3 ***Performance Metrics***

It is important to note that following the processes that we have advised being cross-validated, some way to measure their efficacy is going to be necessary. To evaluate the performance of our categorization systems, we used a number of established metrics in this research. A machine learning model's predictive ability may be evaluated using performance indicators such as accuracy, ROC-curve, precision, recall, and f1-score [20].

Precision: Precision is defined as the quantity of correct diagnoses divided basically by the sum of all assessments, correct and wrong.

Recall: Recall is equal to the sum of true positives divided by the sum of true positives and false negatives.

The F1-score is a calculated average of recall and accuracy.

Accuracy: Split the total number of prediction estimations by the total amount of correct predictions.

Machine learning techniques are evaluated using the following metrics: F1-score value, recall, accuracy, and precision (Sokolova et al., 2006) [21].

Our confusion matrix assessed accuracy, F1-score, recall, and precision for every hierarchical arrangement used. A representation of the effectiveness of the approach is the ML confusion matrix. Both throughput and user input datasets are affected. (Yağanoğlu and Köse, 2018) [22].

5. RESULTS AND DISCUSSIONS

When it came to making prognoses for lung cancer, we used a wide variety of classification strategies in our study. Lung cancer prediction using 5 Boosting machine learning classification models which are part of the Scikit-learn package in A model that forecasts for the detection of lung cancer in patients is being developed using Python, an application of the programming language. The code uses 5 different machine learning algorithms, including XGBoost, Adaboost, Catboost, LGBM and gradient boosting classifier, to predict the likelihood of lung cancer based on a range of variables. The code makes use of a dataset that has a number of columns, including things like gender, age, and smoking, YellowFingers (YFN), anxiety (ANX), PeerPressure (PPR), ChronicDisease (CDG), fatigue (FTG), allergy (ALG), wheezing (WHZ), Alcohol_Consuming (ALC), coughing (CHG), Shortness_of_Breath (SBG), SwallowingDifficulty (SWD), ChestPain (CHP), and LungCancer. The models used for prediction are capable of precisely measuring a patient's chance of acquiring lung cancer by analysing these data and use machine learning algorithms to detect associations and patterns. Here ADASYN is used as Data balancing purpose and Cross validation with 10 K-Fold has been used.

Regarding the purpose of conducting an evaluation of the findings, it is necessary to carry out four different tasks.

1. The importation libraries and datasets, finding correlations, and completing the first task

2. Experimentation with Data Analysis (EDA), which is the second task.
3. The gathering of data for the assessment of the Boosting Based Machine learning model with k-Fold CV, the third task
4. Performance Evaluation and Analysing final results Summary is the fourth task.

	YFN	ANX	PPR	CDG	FTG	ALG	WHZ	ALC	CHG	SWD	CHP	LCR
YFN	1	0.56	0.31	0.02	-0.1	-0.15	-0.06	-0.27	0.02	0.33	-0.1	0.19
ANX	0.56	1	0.21	-0.01	-0.18	-0.16	-0.17	-0.15	-0.22	0.48	-0.12	0.14
PPR	0.31	0.21	1	0.04	0.09	-0.07	-0.04	-0.13	-0.07	0.33	-0.07	0.2
CDG	0.02	-0.01	0.04	1	-0.1	0.13	-0.04	0.01	-0.16	0.07	-0.05	0.14
FTG	-0.1	-0.18	0.09	-0.1	1	-0	0.15	-0.18	0.15	-0.12	0.01	0.16
ALG	-0.15	-0.16	-0.07	0.13	-0	1	0.17	0.38	0.21	-0.04	0.25	0.33
WHZ	-0.06	-0.17	-0.04	-0.04	0.15	0.17	1	0.26	0.35	0.11	0.14	0.25
ALC	-0.27	-0.15	-0.13	0.01	-0.18	0.38	0.26	1	0.2	-0	0.31	0.29
CHG	0.02	-0.22	-0.07	-0.16	0.15	0.21	0.35	0.2	1	-0.14	0.08	0.25
SWD	0.33	0.48	0.33	0.07	-0.12	-0.04	0.11	-0	-0.14	1	0.1	0.27
CHP	-0.1	-0.12	-0.07	-0.05	0.01	0.25	0.14	0.31	0.08	0.1	1	0.19

Table-1: Correlation Values

Task 1 - Importing libraries and dataset and finding Correlation:
The GENDER and LUNG CANCER characteristics in this dataset are of a particular data type. Now we can use sklearn's Label Encoder to turn them into numbers. A utility class called Label Encoder may be used to normalize labels such that they are limited to quantities within zero and n_classes-1. If the labels are hashable and similar, it can also convert them to quantitative ones. Each of the other characteristic may also be set to YES=1 and NO=0. The data for corelation values is as shown in Table 1 and Corelation representation is as shown in Figure 2.

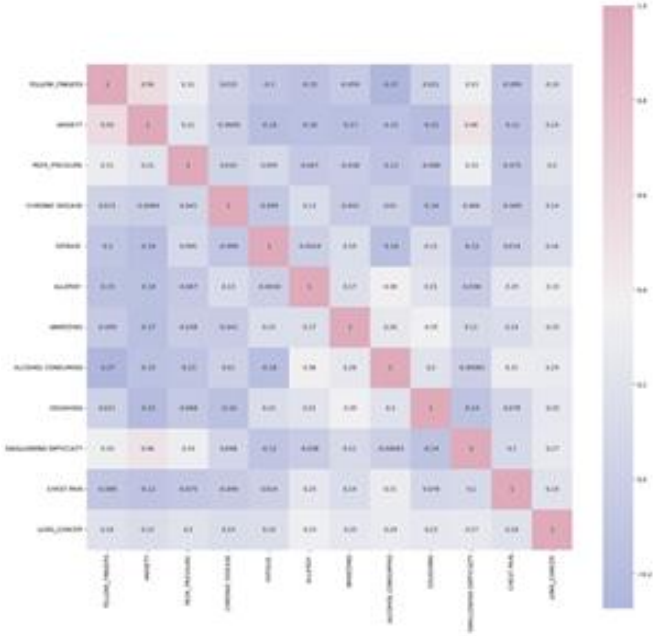

Figure 2: Corelation Values

The correlation matrix reveals that ANXIETY and YELLOW_FINGERS are associated to a greater extent than fifty percent, as was shown in the resulting correlation plot of all characteristics, which can be seen in figure 2. So, lets create a new feature combining them.

Task 2 - Exploratory Data Analysis (EDA):

We will be replacing them with Nan and making provisions for filling in the values that are missing since an amount of zero might be seen as a missing value for the above characteristics. Therefore, there are some NULL values. We shall use a plan to fill in the blanks. We benefit from the fact that the majority of the characteristics follow a roughly gaussian distribution.

Some points to note here are Distribution of target variables.

Let's check the distribution of Target variable. That is, Target Distribution is imbalanced. Before using the approach, we shall address the disparity. To further comprehend the relationship between the independent characteristics and the dependent variable, let's do some data representations.

a). Bar plot for Gender and Lung Cancer

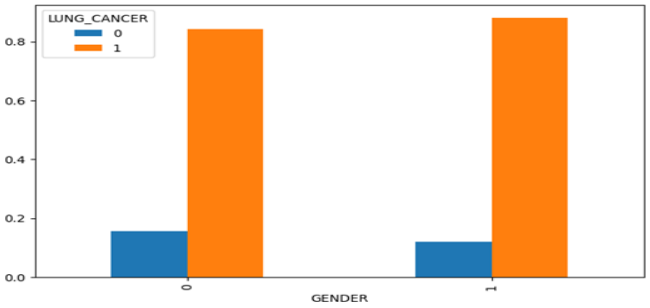

Figure 2a: Bar plot for Gender and Lung Cancer

As a result of this plot (Figure 2a), we are able to draw the conclusion that, in general, the likelihood of getting lung cancer is higher for males.

b). Plot for Age and Lung Cancer

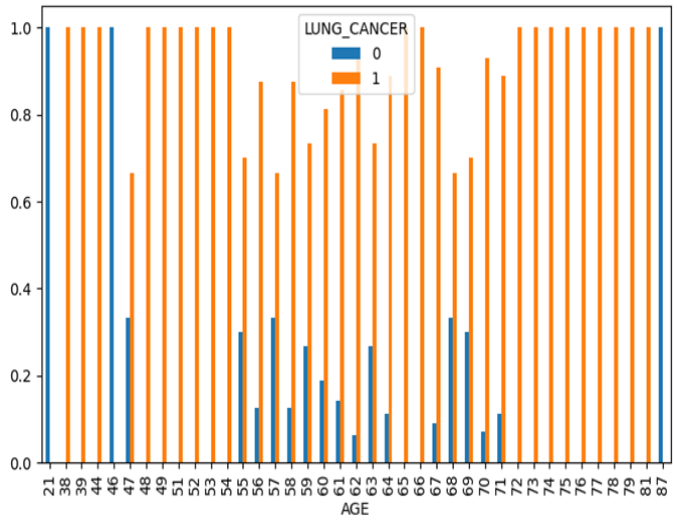

Figure 2b: Plot for Age and Lung Cancer

From this plot (Figure 2b), It is revealed to us that the likelihood of having Lung Cancer almost same for all ages (more nearly after 38 Years).

c). Plot for Smoking and Lung Cancer: Plot shows that chances of lung cancer is more for Smoking persons. (Figure 2c).

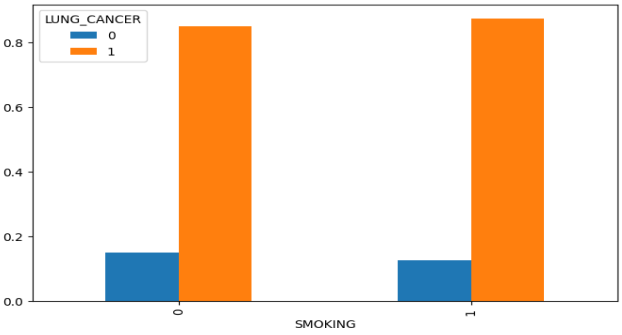

Figure 2c: Plot for Smoking and Lung Cancer

d). Plot for Yellow Finger and Lung Cancer: Plot shows that chances of lung cancer is more for Yellow Fingers. (Figure 2d).

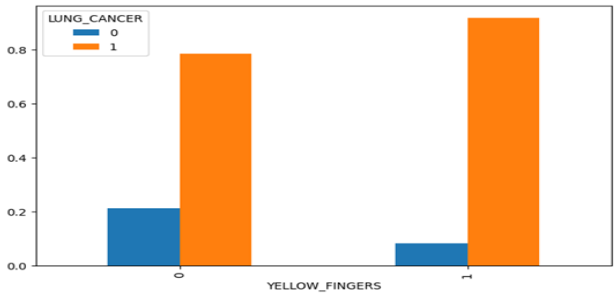

Figure2d: Plot for Yellow Finger and Lung Cancer:

e). Plot for Anxiety and Lung Cancer: Plot shows that chances of lung cancer is more for Anxiety. (Figure 2e).

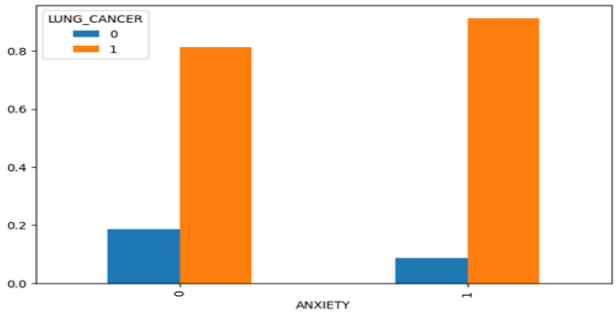

Figure 2e: Plot for Anxiety and Lung Cancer

f). Plot for peer pressure and Lung Cancer: Plot shows that chances of lung cancer is more for peer pressure (Figure 2f)

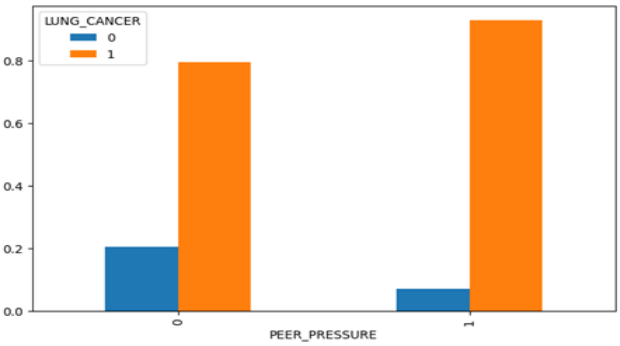

Figure 2f. Plot for peer pressure and Lung Cancer

g). Plot for Chronic Diseases and Lung Cancer: Plot shows that chances of lung cancer are m ore for Chronic Diseases. (Figure 2g)

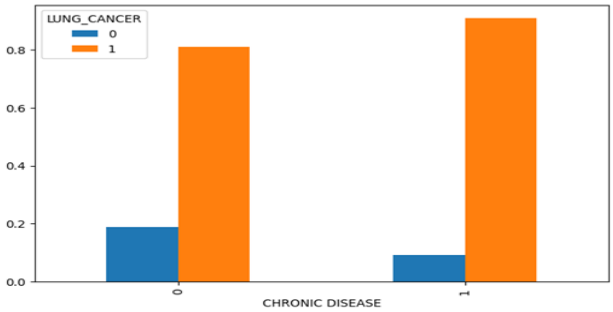

Figure 2g: Plot for Chronic Diseases and Lung Cancer

h) Plot for Fatigue and Lung Cancer: Plot shows that chances of lung cancer are more for Fati gue. (Figure 2h)

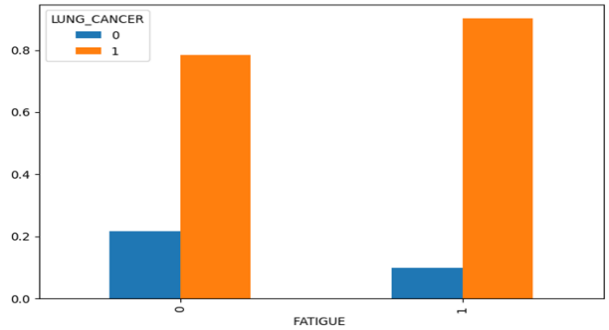

Figure 2h: Plot for Fatigue and Lung Cancer

i) Plot for Allergy and Lung Cancer: Plot shows thatchances of lung cancer are more for Al lergy. (Figure 2i)

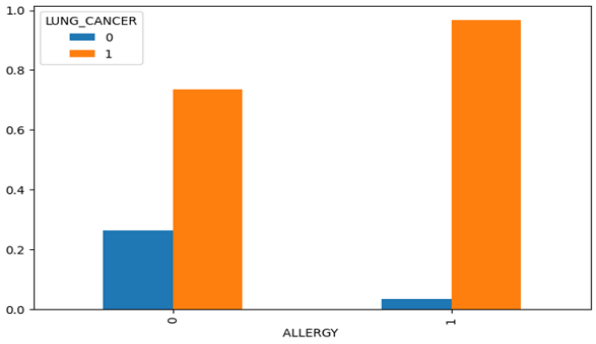

Figure 2i: Plot for Allergy and Lung Cancer

j) Plot for Wheezing and Lung Cancer: Plot shows that chances of lung cancer are more for W heezing. (Figure 2j)

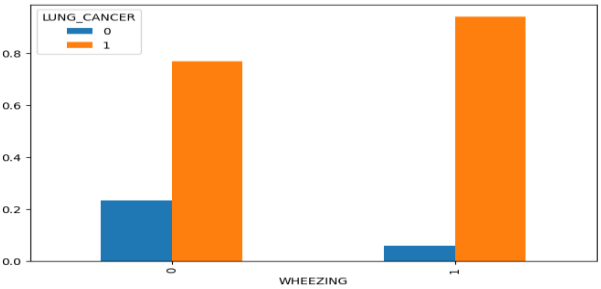

Figure 2j: Plot for Wheezing and Lung Cancer

k) Plot for Alcohol Consumption and Lung Cancer: Plot shows that chances of lung cancer ar e more for Alcohol Consumption. (Figure 2k)

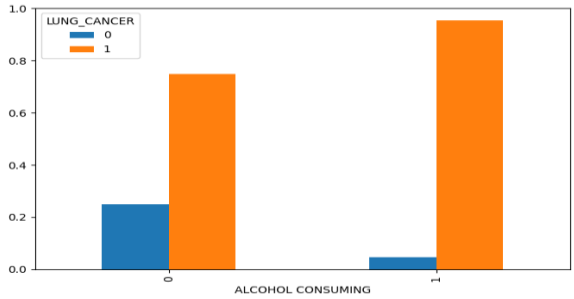

Figure 2k: Plot for Alcohol Consumption and Lung Cancer

l) Plot for Shortness of Breath and Lung Cancer: Plot shows that chances of lung cancer are m ore for Shortness of Breath. (Figure 2l)

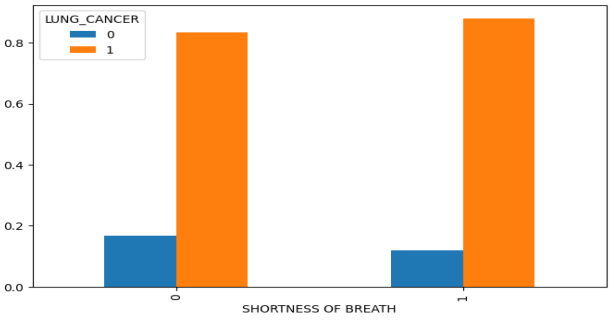

Figure 2l: Plot for Chronic Diseases and Lung Cancer

m) Plot for Chest Pain and Lung Cancer: Plot shows that chances of lung cancer are more for Chest Pain. (Figure 2m)

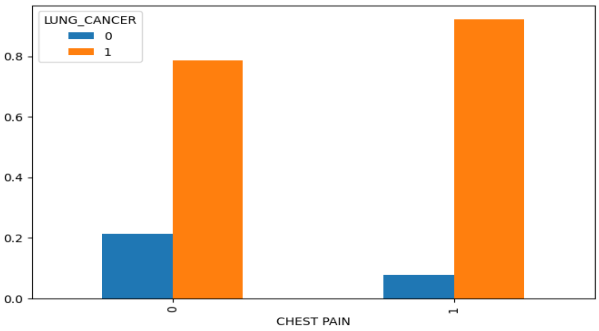

Figure 2m: Plot for Chest pain Diseases and Lung Cancer

n) Plot for Coughing and Lung Cancer: Plot shows that chances of lung cancer are more for C oughing. (Figure 2n)

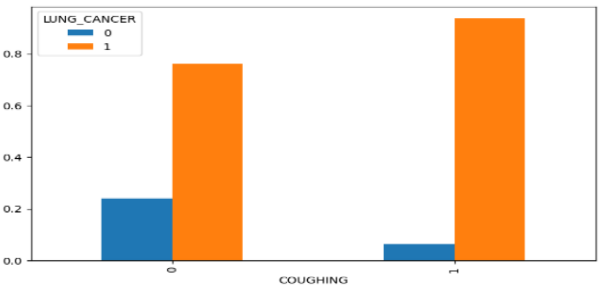

Figure 2n: Plot for Coughing and Lung Cancer

o) Plot for Swallowing Difficulty and Lung Cancer: Plot shows that chances of lung cancer ar e more for Swallowing Difficulty. (Figure2o)

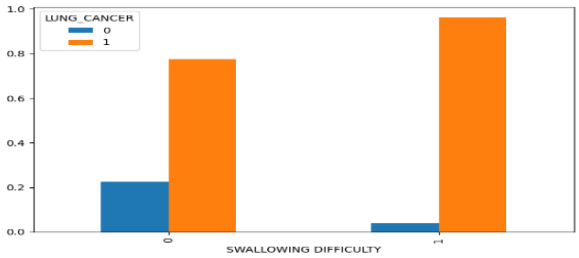

Figure 2o: Plot for Swallowing Difficulty and Lung Cancer

Task-3: The preparation of data for the assessment of different Boosting based Machine Learning models:

The task of checking the outliers is completed in the following steps for Performing CV for Different Models.

a. GB Classifier model. (Table 3a)

	Gradient Boost Classifier			
	Precesion	**Recall**	**F1-Score**	**Support**
Class-0	0.98	0.98	0.98	64
Class-1	0.96	0.98	0.98	56
Accuracy			0.98	120
Macro Avg	0.97	0.98	0.97	120
Wtg.Avg	0.98	0.98	0.98	120

b. Running the XGB Classifier algorithm through cross validation. (Table 3b)

	XGBoost			
	Precesion	**Recall**	**F1-Score**	**Support**
Class-0	0.98	0.97	0.98	64
Class-1	0.96	0.98	0.97	56
Accuracy			0.97	120
Macro Avg	0.97	0.98	0.97	120
Wtg.Avg	0.98	0.97	0.98	120

c. Running the AdaBoost Classifier model through cross validation. (Table 3c)

	AdaBoost			
	Precesion	**Recall**	**F1-Score**	**Support**
Class-0	0.96	1	0.98	64
Class-1	1	0.95	0.97	56
Accuracy			0.97	120
Macro Avg	0.98	0.97	0.97	120
Wtg.Avg	0.98	0.97	0.97	120

d. Cross-validating the LightBGM Classifier model. (Table 3d)

	LGBM Classifier			
	Precesion	**Recall_**	**F1Score**	**Support_**
Class-0	0.89	0.88	0.88	64
Class-1	0.86	0.88	0.87	56
Accuracy	0.88			120
Macro_Avg	0.87	0.88	0.87	120
Wtg_Avg	0.88	0.88	0.88	120

e. *The CATBOOST Classifier model is being cross-validated.*
(Table 3e)

	CATBoost			
	precision	**Recall**	**F1Score**	**Support**
Class-0	0.91	0.98	0.95	64
Class-1	0.98	0.89	0.93	56
Accuracy	0.94			120
Macro_Avg	0.95	0.94	0.94	120
Wtg_Avg	0.94	0.94	0.94	120

Task 4: Final scores of all models

A computer-assisted lung cancer detection method was developed using the dataset as its basis. Which is built using Five distinct machine learning (ML) classifiers including GB, XGBoost, LBGM, ADABOOST, and CATBOOST, have been presented in this research. Before using classification strategies, we have first pre-processed each unique data point included inside the dataset. Table 4 and figure 4 shows Accuracy, Precision, Recall, F1-Score, for all Boosting based ML Methods in order to assist the rapid selection of the optimum model while taking into consideration all of the scores simultaneously. For this particular example, GB Classifier yielded the highest possible Accuracy. Due to the data's imbalance, the ADASYN oversampling approach was used to equalize the number of positive and negative cases. Various models have been tested here. Outcomes might be significantly improved with more meticulous tuning, development of features, and cross-validation techniques.

Table 4: Comparison of Boosting algorithms-based ML Models for accuracy and other parameters.

Boosting Based Machine Learning Classifiers				
	precision	**Recall_**	**F1_Score**	**Accuracy_**
GB	0.98	0.98	0.98	0.98
LGBM	0.87	0.88	0.87	0.88
CatBoost	0.95	0.94	0.94	0.94
AdaBoost	0.98	0.97	0.97	0.97
XGB	0.97	0.98	0.97	0.97

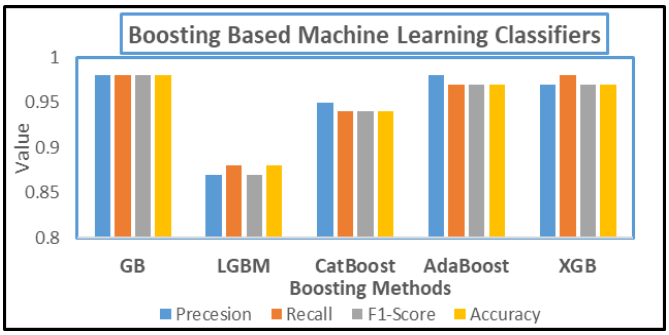

Figure 4: Performance Evaluation of different Boosting Algorithms

Table 5: Comparison of Boosting algorithms-based ML Models for Average Scores in %.

Boosting Based Machine Learning Classifiers			
	Average Accuracy %	**Cross Validation Score %**	**ROC AUC Score %**

GB	**95.61**	**97.14**	**98.33**
LGBM	86.99	88.01	87.5
CatBoost	90.76	91.57	93.86
AdaBoost	94.14	97.13	97.32
XGB	95.18	97.14	97.54

Final Average Scores of boosting Algorithms as shown in Table 5 and Figure 5 shows that Gradient Boosting Algorithms shows highest values of average accuracy of 95.61%, Cross validation score of 97.14%, ROC AUC Score of 98.33 % and LGBM algorithm shows Lowest values of average accuracy of 86.99%, Cross validation score of 88.01%, ROC AUC Score of 87.5 %. Thus, GB outperforms with respect to other methods

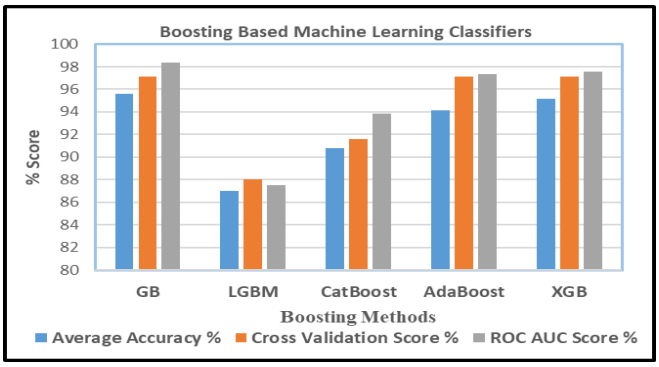

Figure 5: Performance Evaluation of different Boosting Algorithms for Average Scores in %

5. CONCLUSION

It is possible that ML, which is a subset of AI, may revolutionize the process of predicting lung cancer risk and detecting it at an early stage. Successfully managing Lung Cancer requires early detection. We made preparations, conducted evaluations, and carried out Prediction of Lung Cancer Using a Number of Machine Learning (ML) Techniques based on Boosting methods or classifiers and output evaluation was carried out for the purpose of identifying the most effective classifier with the highest degree of accuracy. The properties of the data set were gathered and analysed by us by using boosting algorithms-based ML classification methods in this research work to attain high accuracy. GB algorithm outperform with respect to other boosting algorithms-based classification methods. Age and Lung Cancer are unrelated, despite the fact that there is scientific proof to the contrary. Machine learning (ML) powered by artificial intelligence (AI) classifiers have transformed risk assessment for Lung Cancer in its earliest stages. We used boosting algorithms-based machine learning classification algorithms and Lung Cancer risk variables for the purpose of predicting Lung Cancer early in our work. five classification algorithms: GB, XGB, CatBoost, Adaboost, LGBM were tested on the Lung Cancer dataset. GB beat other machine learning (ML) methods while dealing with Lung Cancer's primary stages detection by over 95.61 % and LGBM Shows lowest Average Accuracy of 86.99 %. Our research work can basically properly predict Lung Cancer but has limits. Due to the small size of the study's sample, it was impossible to determine whether or not the results were statistically significant. For the purpose of improving our ability to categorize illnesses, we would want to collect more worldwide data and Apply hybrid combinations of different boosting-based Machine Learning classifiers with different Cross Validations Methods to enhance accuracy.

6. REFERENCES

1. I. D. Mienye, Y. Sun and Z. Wang, "Improved predictive sparse decomposition method with densenet for prediction of lung cancer", Int. J. Comput., vol. 19, no. 4, pp. 533-541, Dec. 2020.
2. I. D. Mienye, G. Obaido, K. Aruleba and O. A. Dada, "Enhanced prediction of chronic kidney disease using feature selection and boosted classifiers" in Intelligent Systems Design and Applications, Cham, Switzerland, pp. 527-537, 2022.
3. Y. Sun, Z. Li, X. Li and J. Zhang, "Classifier selection and ensemble model for multi-class imbalance learning in education grants prediction", Appl. Artif. Intell., vol. 35, no. 4, pp. 290-303, Mar. 2021.
4. M. Kearns and L. G. Valiant, "Crytographic limitations on learning Boolean formulae and finite automata", Proc. 21st Annu. ACM Symp. Theory Comput. (STOC), pp. 433-444, 1989.
5. R. E. Schapire, "The strength of weak learnability", Mach. Learn., vol. 5, no. 2, pp. 197-227, 1990.
6. Y. Freund and R. E. Schapire, "A short introduction to boosting", Proc. 16th Int. Joint Conf. Artif. Intell., pp. 1401-1406, 1999.
7. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system", Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, pp. 785-794, Aug. 2016.
8. A. Shimazaki, D. Ueda, A. Choppin et al., "Deep learning-based algorithm for lung cancer detection on chest radiographs using the segmentation method," *Scientific Reports*, vol. 12, no. 1, Article ID 727, 2022.
9. S. Bhatia, Y. Sinha, and L. Goel, "Lung cancer detection: a deep learning approach," in *Soft Computing for Problem Solving: SocProS 2017, vol. 2*, pp. 699–705, Springer, 2019.
10. M. A. Talukder, M. M. Islam, M. A. Uddin, A. Akhter, K. F. Hasan, and M. A. Moni, "Machine learning-based lung and colon cancer detection using deep feature extraction and ensemble learning," *Expert Systems with Applications*, vol. 205, Article ID 117695, 2022.
11. Suresh Kumar Tummala, Phaneendra Babu Bobba & Kosaraju Satyanarayana (2022) SEM & EDAX analysis of super capacitor, Advances in Materials and Processing Technologies, 8:sup4, 2398-2409,
12. J.Marcello, F.Marques, and F.Eugenio, "Evaluation of thresholding techniques applied to ocean graphic remote sensing imagery",SPIE, 5573,pp. 96-103, (2004).
13. Vesna Zeljkovic, MilenaBojic, "Automatic Detection of Abnormalities in Lung Radiographs caused by plan cellular Lung Cancer", IEEE, (2011)
14. Marjolein A. Heuvelmans et," A Survey on Lung Cancer Prediction" Research Gate publication /377381070.
15. T. Chen and C. Guestrin, "Xgboost: Reliable large-scale tree boosting system," in Proceedings of the 22nd SIGKDD Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 2015, pp. 13–17.
16. Freund, Y., Schapire, R.E.: "A decision-theoretic generalization of on-line learning and an application to boosting". Journal of Computer and System Sciences 55(1), 119–139 (1997)
17. Friedman, J. (2001). "Greedy boosting approximation: a gradient boosting machine". Ann. Stat. 29, 1189–1232. doi: 10.1214/aos/1013203451
18. https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

19. https://proceedings.neurips.cc/paper/2018/file/14491b756b3a51daac41c24863285549-Paper.pdf
20. D. M. W. Powers, "Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation," Journal of Machine Learning Technologies, vol. 2, no. 1, pp. 37–63, 2011.
21. Sokolova M., Japkowicz N., Szpakowicz S., (2006), "Beyond Accuracy, F-score and ROC: a Family of Discriminant Measures for Performance Evaluation", American Association for Artificial Intelligence (www.aaai.org).
22. Yağanoğlu, M., & Köse, C., (2018)," Real-time detection of important sounds with a wearable vibration-based device for hearing-impaired people". Electronics, 7(4),