Journal of Neonatal Surgery
ISSN(Online): 2226-0439
Vol. 14, Issue 5s (2025)
https://www.jneonatalsurg.com

OPEN ACCESS

Prediction and Segmentation of Heart Disease Boosting-Based Machine Learning Algorithms

Ashok Kumar[1], Deepika Dhamija[2], Vikrant Chole[3], Jhankar Moolchandani[*4], Rahul Kumar[*5], Umang Garg[6]

[1,3,4,5]Department of Computer Science and Engineering, Amity School of Engineering and Technology, Amity University Madhya Pradesh, Gwalior, India.

Email ID: ashok_gangwar@rediffmail.com

Email ID: vikrantchole@gmail.com

[2]Centre for distance and Online Education Manipal University Jaipur, India.

deepika.dhamija@jaipur.manipal.edu

[6]Computer Science and Engineering, School of Computing, MITADT University, Pune, India.

Email ID: umangarg@gmail.com

***Corresponding Author**

Email ID: Jmoolchandani@gwa.amity.edu, Email ID: rahulkumar1680@gmail.com,

Cite this paper as: Ashok Kumar, Deepika Dhamija, Vikrant Chole, Jhankar Moolchandani, Rahul Kumar, Umang Garg, (2025) Prediction and Segmentation of Heart Disease Boosting-Based Machine Learning Algorithms. *Journal of Neonatal Surgery*, 14 (5s), 324-334.

ABSTRACT

Recent advances in imaging and sequencing technologies have led to significant advancements in clinical research on lung cancer. However, the amount of information that the human brain can properly digest and utilize is limited. Lung cancer has been extensively detailed by integrating and analyzing this vast and complex amount of data from a variety of perspectives. Machine learning-based technologies are essential to this process. This study tests multiple Boosting algorithm models on a lung cancer dataset to determine a particular lung cancer disease prediction. The aim of this work is to determine the best cross-validation methods and boosting algorithms to enhance performance in lung disease predicting. The effectiveness of the method is evaluated using a number of performance metrics, such as recall, accuracy, precision, F-score, ROC AUC score, and cross validation score. The famous Lung Cancer Dataset is used in this academic paper to test a number of machine learning classification techniques based on boosting algorithms, including Gradient Boost (GB), Extended Boost - XGBOOST (XGB), Adaptive Boost (ADABOOST), Categorized Boost (CATBOOST), and Light Gradient Boost (LGBM). many Kfold cross-validation techniques. The impact of the ADASYN as a data balancing approach on the precision of lung cancer prediction employing algorithms is investigated through hybrid combinations of cross validation and boosting procedures. This study presents a hybrid approach that could accurately predict the incidence of lung cancer. This study discovered that a hybrid integration of the Cross-validation approach with data balancing and the Boosting based ML Models built utilizing machine learning-based modeling category worked well to produce more accurate predictions regarding lung cancer.

Keywords: *Lung Cancer prediction; Gradient Boost, XGBOOST, ADABOOST, CATBOOST, Cross validations, Data Balancing.*

1. INTRODUCTION

Lung cancer is the leading cause of death from cancer worldwide [1]. At 18% of all cancer-related deaths, it is the leading cause of death among all types of cancer. Smoking has either reached its peak or is still the leading cause of lung cancer in many countries. This suggests a rise in lung cancer incidence in the ensuing decades [2]. With timely and accurate diagnosis, lung cancer patients may have a significantly better prognosis [3]. After receiving a lung cancer diagnosis, between 10% and 20% of patients live for five years. Common medical procedures for early detection that improve patient survival rates include computed tomography (CT) and magnetic resonance imaging (MRI) [4]. The process of repeatedly applying the underlying learning algorithm to updated input data is known as "boosting" [5]. Boosting algorithms use input data to train a weak

learner, calculate the learner's predictions, first select incorrectly identified training samples, and then use an updated training dataset that includes the instances that were incorrectly classified in the previous training cycle to teach the next weak learner [6].

Lung cancer ranks highest in terms of incidence and mortality worldwide when compared to other cancer forms. Approximately 2.20 million new cases of lung cancer are discovered each year [7], and 75% of those people pass away from the disease within five years of being diagnosed [8]. High intra-tumor heterogeneity (ITH) and the complex nature of cancer cells, which leads to drug resistance, make cancer therapy more challenging [9]. Over the past few decades, numerous large-scale collaborative cancer research has been made possible by the ever-increasing technical capabilities in cancer study, leading to the production of numerous clinical, medical imaging, and genetic databases [10–11]. These datasets aid in diagnosis, treatment, and responses to clinical outcomes by allowing researchers to look at full cases of lung cancer [12]. This study's main objectives were to:

- To ascertain whether publicly accessible datasets exist in the field of lung cancer research. The GB, XGBOOST, ADABOOST, CATBOOST, and LGBM are a few examples of boosting-based machine learning models whose performance will be thoroughly evaluated.
- To examine the impact of applying the data balancing technique and carry out a thorough performance assessment of data balancing and cross-validation methodologies.
- To assess how well boosting algorithms and cross-validation approaches work together with data balance to improve the accuracy of lung cancer prediction.
- To use performance measures to evaluate the efficacy of early lung cancer detection.

The rest of the research paper is offered after this portion of the introduction and the structure, and in Part II, we examine the relevant research literature. We give a high-level overview of the methods and strategies we employed in our research in the third section. However, whereas Section IV gives details on the recommended approach and performance measures, Section V presents the results of the studies. The results and their consequences are highlighted in the final section, known as VI.

2. LITERATURE SURVEY

Lung cancer is one of the leading causes of death worldwide and has the highest prevalence. Throughout the past few decades, lung cancer has been one of the most common cancer cases worldwide. Approximately 2.1 million new cases of lung cancer were discovered worldwide in 2018, according to [13]. This represents 12% of all cancer cases reported worldwide. It is important to highlight that the overall five-year survival rate for those with lung malignancies is only 18%. However, the percentage of individuals who survive lung cancer could increase to about 55% if the disease can be identified sooner. Patients who are diagnosed with lung cancer at an early stage have been shown to have a survival rate of up to forty percent over a period of five years provided they get the right therapy [14]. It is unfortunate that more than seventy percent of patients get a diagnosis after their tumor has already reached an advanced stage, and most of these instances are not candidates for surgical intervention. There is a connection between this and the fact that the diagnostic procedures that are now in use are not precise or accurate enough. CT-guided transthoracic aspiration biopsy is now considered the gold standard for identifying lung cancer. However, this procedure is not only costly but also entails the risk of respiratory complications such as pneumothorax, pulmonary embolism, and substantial trauma. Consequently, it is not appropriate for most patients. In addition to a breathing test and blood tumor biomarkers, there are a great number of alternative diagnostic tests that may be used for lung cancer monitoring; however, each of these approaches has its own set of restrictions [15]. As a result, it is essential to discover useful diagnostic biomarkers for the case related to lung cancer, particularly for lung cancer in its early stages.

Puneet et al. [16] developed a framework for predicting lung cancer through the application of machine learning strategies that were founded on regular blood indicators obtained from their research. To forecasting the outcomes, they used several different classifiers, including XGBoost, Grid SearchCV, LR, SVM, GNB, NB, DT, and KNN classifiers employed using K-fold 10-cross-validation. It was Lanzhou University that was responsible for collecting the dataset, and it included a total of 277 cases. When compared with various other classifiers, the authors discovered that XGBoost fared much better in terms of accuracy (92.16%), recall (96.97%), and area under the curve (95%).

The detection and prediction of lung cancer was accomplished by Faisal et al. [17] using several machine learning and ensemble learning techniques. MLP, NN, NB, SVM, Majority Voting (Hard and Soft), GB, and RF by K-fold 10 cross-validation were the methods that the authors used to observe the computational capabilities of the models. An accuracy of 90%, precision of 87.82%, recall of 83.71%, and F1-score of 85.71% were reached by the Gradient Boosted Tree (Ensemble Learning approach), according to the researchers' observations. The methodology surpassed all other distinct classifiers. This dataset was obtained from the database operated by the University of California, Irvine (UCI), and it includes 32 occurrences and 57 characteristics.

In their study, Safiyari et al. [18] utilized a variety of basic ensemble methodologies for learning, including those of Bagging, Dagging, AdaBoost, MultiBoosting, and Random SubSpace. Additionally, they utilized several other classification techniques, including RIPPER, Decision Stump, SimpleCart, C4.5, SMO, Logistic_Regression, Bayes_Net, and Random_Forest, to predict the survival rate of lung cancer patients. The under-sampling strategy was used by the authors to assess the forecasting model on the Surveillance, Epidemiology, and End Results (SEER) dataset. This dataset includes 643,924 observations and 149 characteristics. The outcomes were analyzed, and it was discovered that the AdaBoost method fared better than other approaches in terms regarding the area under the curve and the accuracy metrics, which were correspondingly 94.9 % and 88.98 %.

Muntasir et al. [19] conducted yet another scientific study about lung cancer. In the present investigation, they scrutinized a number of earlier research that focused on developing modeling systems for the prediction of lung cancer and compared the outcomes of those studies with their own models. Several different methods, including XGBoost, LightGBM, AdaBoost, and bagging collective learning, were developed by them in order to provide predictions about lung cancer. The K-fold (k=10) CV methodology was used in the process of validating the mathematical model. This investigation has a maximum accuracy of 94.42%, which is accomplished by the use of XGBoost.

Patra [20] investigated a number of various classifiers developed using machine learning for the purpose of detecting lung cancer. These classifiers included RBF, KNN, J48, SVM, LR, ANN, NB, and RF. In all, there are 32 occurrences and 57 characteristics that are included in the dataset that was obtained from the "UCI repository." An accuracy of 81.25% was attained by RBF, which was by the researchers considered to be superior than the accuracy obtained through any of the other approaches. Using a number of different machine learning simulations, which includes as DT, LR, Bagging, RF, and AdaBoost, Sim et al. [21] proposed doing research on health-associated quality of life (HRQOL) in the context of predicting mortality from lung cancer over a period of five years. For the purpose of evaluation of the performance of the model, two distinct sets of characteristics were deployed in conjunction with Kfold 5 cross-validations. A comparison was made between the effectiveness of the model with the data collected from 809 lung cancer surgery survivors who participated in the procedure. Based on the proposed results of this research work under considerations, it was determined that AdaBoost had the greatest accuracy, which was 94.8%.

3. PROPOSED MODEL

The preliminary processing phase of the proposed technique begins with the collection of pertinent data. The selected categories are then trained and evaluated on the lung cancer dataset using the well-liked ten-fold cross-validation process. XGBoost, AdaBoost, GB, CatBoost, and LightGBM are a few of these. In order to observe the effect of hybrid combinations of boosting methods and cross validation methods with ADAYN as a data balancing method on the accuracy of lung cancer prediction, the cross-validation techniques of Kfold (KF) are combined with the boosting algorithms with and without ADASYN as a data balancing approach. The results of this investigation show that a hybrid approach can accurately predict lung cancer. According to the results of this study, model-based machine learning classifiers that use boosting algorithms and the K-Fold Cross validation technique with ADASYN as a data balancing method have better accuracy and precision when forecasting lung cancer. The main goal of this research project is to design, implement, and assess the results of the Lung Cancer Prediction study using a variety of machine learning approaches in order to identify the best effective classification algorithm. In keeping with this, we shall briefly review the next stages.

3.1. Dataset Description Phase

For this study, we use the publicly accessible Lung Cancer Data Set, which can be found on the Kaggle website (Dataset, Lung Cancer Data Set). This dataset contains 309 entries and 16 characteristics, one of which is the Lung_Cancer property. The remaining 89 cases are "tested negative," meaning that the patient essentially does not have lung cancer, while the remaining 270 cases are "tested positive," meaning that the patient has lung cancer.

3.2. Data Pre-processing phase

The initial evaluation of the data is valuable. When machine learning algorithms are used to the dataset, this analysis approach ensures accurate predictions and dependable results (Soni et al., 2020) [22]. Although the Indian Lung Cancer dataset has empty values for a few unhelpful features, it does not contain any missing values (NAN) values. While there are no missing values (NAN) in the Indian Lung Cancer dataset, there are zero values for several unhelpful attributes. We include zero values for patients without lung cancer and those with lung cancer to obtain the required average and median values for each column. Diabetes patients and non-diabetic patients both swaps out zero. Validation and training were conducted on 75% of the standardized Lung Cancer Data Set, with the remaining 25% being used for evaluation. Python is used in the framework development process.

3.3. General Introduction of Boosting Methods: GB, CatBoost, XGB, AdaBoost, LGBM

The base learner must first take all of the observations and assign the same weight or level of attention to each one.

Observations with prediction mistakes are given additional weight in the second phase if the initial base learning approach produced any prediction errors. The data is subsequently subjected to the following fundamental learning technique. Proceed to Step 3 and repeat Step 2 until the basic learning method's limit is reached or a higher degree of accuracy is achieved. Last but not least, it improves the framework's ability to produce precise predictions by merging the outputs of weak learners to create a strong learner. In the process of boosting, examples with more errors or those that have been misclassified are given more weight than weak rules.

As many machine learning classifiers as possible are used in ensemble learning to train the model. Bagging is an example of an ensemble learning technique that uses multiple models to segregate dataset subsamples. Boosting is widely used and trains both the method and the model, unlike parallel construction. The model is trained using a simple technique and then restructured to facilitate learning. For ease of understanding, the next method makes use of the modified model. Numerous progressive boosting techniques with unique approaches are covered in this article. We use machine learning to classify our dataset as it becomes available. The study incorporates CatBoost, AdaBoost, XGBoost, GB, and LGBM computations, including factors such as smoking, yellow_fingers, anxiety, peer_pressure, chronic_disease, and fatigue. Allergy, Wheezing, Alcohol_Consumption, Coughing, Shortness_of_Breath, Swallowing_Difficulty, Chest_Pain, Lung Cancer, and two Exploratory Data Analysis characteristics.

3.3.1 XGBoost (XGB): XGBoost is fast, simple, and effective on huge datasets. No parameter optimization or adjustment is needed; thus, it may be utilized immediately after installation. Gradient Boosted decision trees are implemented using XGBoost. Sequential decision trees are constructed in this technique. Weights matter in XGBoost. All independent factors are weighted and supplied into the decision tree to forecast outcomes. Enhance the weight of variables predicted erroneously by the tree and feed them to the second decision tree. Ensembles of classifiers/predictors create a stronger, more accurate model. It solves regression, classification, ranking, and user-defined prediction issues. Gradient Boosted Trees is built easily and efficiently, and the infrastructure offers a viable distributed machine learning environment for scaling tree-boosting algorithms. For rapid parallel tree generation, the classifier is properly configured and fault-tolerant in a distributed environment. A node's billions of data are merged with scale-beyond distributed software samples. [23].

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

3.3.2 AdaBoost (Adaptive Boosting):

It can fit a series of weak learners using a variety of weighted training data. Initially, it makes predictions based on the initial data set, and then it assigns equal importance to each occurrence. When the first learner makes an inaccurate prediction, it assigns more weight to observations that have been forecasted wrongly. This occurs when the prediction is incorrect. Since it is an iterative procedure, it will continue to add learner(s) until a limit about the number of models or accuracy is achieved. We make most of our usage of decision stamps using AdaBoost. However, if the machine learning algorithm permits weight on the training data set, then we can utilize any of the algorithms as the base learner. The AdaBoost technique may be used for solving classification problems as well as regression problems.

One of the most well-known judgment stump-based boosting algorithms is called Adaboost [24]. It is not the case that Adaboost will blindly repeat this method. The estimated weights of many methods are updated in a sequential manner, and each of these techniques contributes to the best accurate estimate possible. Each algorithm does an error calculation. Using the second procedure, weights are updated. 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 bringing measurements up to date and then moving them on to the subsequent stage, the approach makes categorization more accurate. A complicated illustration of a sequential algorithm procedure is as follows:

Consider the labels to be red and blue. Labels are separated by the weak classifier 1, which incorrectly categorizes two blue and one red data. Following that, the subsequent model considers the errors that were made in classification and reduces the weights of the classifications that were correct. Since it incorrectly classifies samples with rising bias and then corrects them while simultaneously reducing bias, the new model picks up new information more quickly than the previous method. The process is repeated in subsequent phases. The use of weak categories is necessary for strong categories. Through the importation of AdaBoostRegressor, regression may be achieved.

3.3.3 Gradient Boost:

It basically trains many models in a sequential fashion using gradient boosting. When utilizing the Gradient Descent approach, every new model developed progressively minimizes the loss function of the whole system, which is represented by the equation y = ax + b + e. The letter e requires particular attention since it represents a term that is incorrect. Over the course of the process of learning, different models were successively fitted to produce a more precise estimation of the response variable. The fundamental concept that underpins this approach is the construction of new base learners that can

achieve the highest possible association between the loss function and the negative gradient of the function that corresponds with the entire ensemble of learners. By including the decision stump into its most recent version to its weighting system, which consisted of one node separated into two leaves, Adaboost was able to improve its performance. Another sequential method, known as gradient boost [25], results in larger trees. This is since the loss is optimized by increasing the number of leaves from 8 to 32. Loss of taxation indicates that for example look at the residual in models that are linear. The amount of loss corresponds to the aggregate of the squares. produced by each of the data scores., and the residual error is equal to the difference between (both of which of) the measured value of y and the projected value of y. Both values are referred to as the difference factor. As to why the square is used. Since the desired outcome represents the difference between what was projected and what really occurred, prediction mistakes are of utmost significance. However, squaring a negative number would result in a little loss, therefore negative numbers are squared. This is the case even if the negative number in question is not zero. In a nutshell, the subsequent method is provided with a collection of residual values, which are then reduced to make them suitable for transmission to the subsequent technique [26].

3.3.4 LightBoost (LGBM)

A technique for tree-based learning is used by the Light GBM The structure which is a gradient boosting framework. What sets it different from other algorithms that relies on trees? Light GBM develops trees in a vertical direction, while other algorithms develop trees in a direction that is horizontal. This means that Light GBM grows trees in a leaf-wise direction, whereas other algorithms develop trees in a level-wise direction. "Light Gradient Boosting Machine" (abbreviated as "LGBM") is a decision tree-based algorithm that was first developed by Microsoft in the year 2017. When compared to earlier methods, this one separates the tree according to the leaves, which means that it is possible to find and deactivate enemy soldiers with exact precision. The LGBM technique is an extremely effective method for minimizing errors while simultaneously maximizing accuracy and speed. The customized approach may be used to split quantitative data; however, in order to do so, an integer number as well, such as an index, must be used in lieu of the column's text name. LightGBM is an ensemble learning framework, more precisely a gradient boosting approach, that involves the successive addition of weak learners in a gradient descent manner. This results in the construction of a strong learner. Memory use and training time are both optimized as a result.

3.3.5 CatBoost

In the year 2017, Yandex developed CatBoost. The One-HotEncoding algorithm, which converts all category traits to numerical values, is the source of the category Boosting algorithm [27]. You might also put the indices value in place of the column name that is not being used. The numerals that are lacking are accommodated. It functions better than XGBoost. However, in contrast to other boosting methods, Catboost makes use of symmetric networks that have the same split in nodes throughout all levels. When training a model, Following the calculation for each of data point, the residual error that was found, XGBoost and LGBM engage in the process of training the predictive model to a remainder target value. Over the course of multiple repetitions, it learns and reduces the amount of residual error in order to accomplish the desired result. Because this method is applied to each individual data point, it has the potential to reduce generalization and lead to overfitting from occurring the number of residuals for each data point will be generated by Catboost via the application of the model that it has trained with to several data points that came before it. Each data point produces its own distinct set of residual data. As many times as these data are evaluated, the general framework is trained to do the evaluations. Because several models will be put into effect, this computing cost is both expensive and exhausting. A more organized increase is completed more quickly. Instead, then beginning with the residual of each individual data point (n+1), sequential boosting begins with the residual of all the data points. To calculate n+2, use the n+1 formula.

3.4 K-fold cross-validation

It is necessary to make use of cross-validation methods in order to guarantee that the model is successfully trained on the data that is supplied without a significant amount of noise. This methodology of statistical estimation is used for the purpose of evaluating the effectiveness of machine learning models [28,29]. In K-Fold cross validation technique, the dataset is folded into k equal-sized folds using this approach. It is called k-fold because it has k pieces, which may be any number like 3,4,5, etc. Validation uses one-fold while model training uses K-1 folds. This is done k times to utilize every fold once as a validation set and other left-outs as training sets (Figure 1).

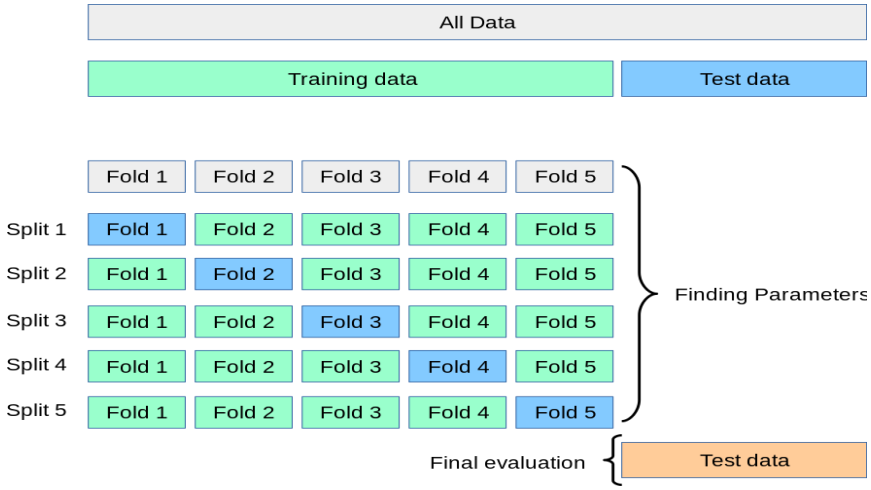

Fig. 1: K-fold cross-validation

The graphic above displays 5 folds and 5 iterations. One-fold is the test/validation set and k-1 sets (4 sets) are the train set in each iteration. Take the k-models validation data accuracy to obtain the final accuracy. For unbalanced datasets, this validation method will not train the model successfully due to the appropriate ratio of each class's data.

4. EXPERIMENTAL SETUP AND CONFIGURATION

Initially, we pre-processed the dataset that was pertaining to lung cancer in our study endeavour. Following preprocessing, the dataset was divided into train and test sets using a technique called tenfold cross-validation (2xCV). After that, the procedures that are indicated are used on the training set as a diagnostic instrument for lung cancer mellitus in its early phases. In conclusion, evaluation measures are used to compare efficacy on the subject of the test. These historical periods are discussed in a concise manner in the following part.

The procedure of resolving missing values in the data that had been pre-processed was an important component of the data pre-processing that enabled us to be successful in achieving our research objective. As an example, the prognosis of lung cancer using machine learning and deep learning is not appropriate for usage with minimum values of the characteristics. To quantify nominal attribute values, such as "male" and "female" in the Gender category, "yes=1" and "no=0" in the other attributes category, and "positive" and "negative" in the class category, namely Lung Cancer, we give a 1 to "yes" and a 0 to "no." This allows us to determine the extent to which these characteristics are significant. One thing that should be brought to your attention is that after we have completed the process of cross-validating our suggested methods, we will need to find a way to evaluate how well they worked. Within the scope of our investigation, we used a wide range of established criteria for evaluating the effectiveness of categorization systems in order to evaluate the outcomes of our several research. In essence, performance measurements like as precision, recall, f1-score, ROC-curve, and accuracy are used in order to ascertain the amount of predictive performance. To find the precision, divide the number of right diagnoses by the number of wrong diagnoses. "F1-score" is the geometric mean of "accuracy" and "recall." To Find Accuracy, divide the number of correct predictions by the total number of forecasts. To rate machine learning methods, we use accuracy, precision, recall, and the F1score number. For each hierarchy order we implemented, our confusion matrix checked the F1-score, memory, accuracy, and precision. The machine learning uncertainty grid displays how well a program works. There are several steps included in the processing of the data.

1. Importing libraries and dataset for Lung cancer prediction: Lung cancer data under consideration is Referred for which Lung cancer is to be predicted and performance enhancement model is to be prepared
2. Finding Corelations for the different features: Different important attributes or features are noted down for the Data set under considerations
3. Exploratory Data Analysis (EDA): EDA's main goal is to help people look at data before they make any decisions. It can help find clear mistakes, understand trends in the data better, find outliers or events that don't fit the pattern, and find interesting connections between the factors.
4. Data Preparation for model evaluation: The technique of making raw data for machine learning models is called data pre-processing. There you have it: the fundamental step in making a machine-learning model. This part of data science is the hardest and takes the most time. In machine learning systems, data needs to be pre-processed to make it less complicated.

5. Adopt Cross Validation Technique: Cross-validation is a way to see how well a machine learning model can guess what new data will be. Plus, it shows issues like overfitting or selection bias and lets you know how the predictive algorithm will work with a different set of data.
6. Data Balancing for model Evaluation for all Boosting Algorithms: Predictive modelling is hard when datasets aren't balanced, but this is a regular phenomenon that we expect to see because the real world is full of cases that aren't balanced. When you balance a dataset, it's easier to train a model because the model doesn't become biased regarding one class.
7. Evaluate the performance of all boosting based algorithms and K-Fold Cross Validation technique: Different Boosting Algorithms used are Gradient Boost (GB), Extended Boost (XGB), LightGBM, Categorical Boost (CatBoost) and Adaptive Boost (AdaBoost) along with K-Fold Cross validation Technique. Different performance metrics used are Precision, Recall, F1-Score, Accuracy, Cross Validation Score, ROC AUC Score
8. Find Final Results Summary and Compare the results for Boosting based algorithms and Cross validation technique with and without data balancing: Effect of Boosting Method and Cross validation methods on performance on Lung cancer detection Accuracy is observed
9. Conclude the Results: The Accuracy is evaluated for Highest value of Hybrid combination of Boosting algorithm and Cross validation Technique and ADASYN as Data Balancing methodology. System block diagram with stepwise implementation is as shown in figure 2.

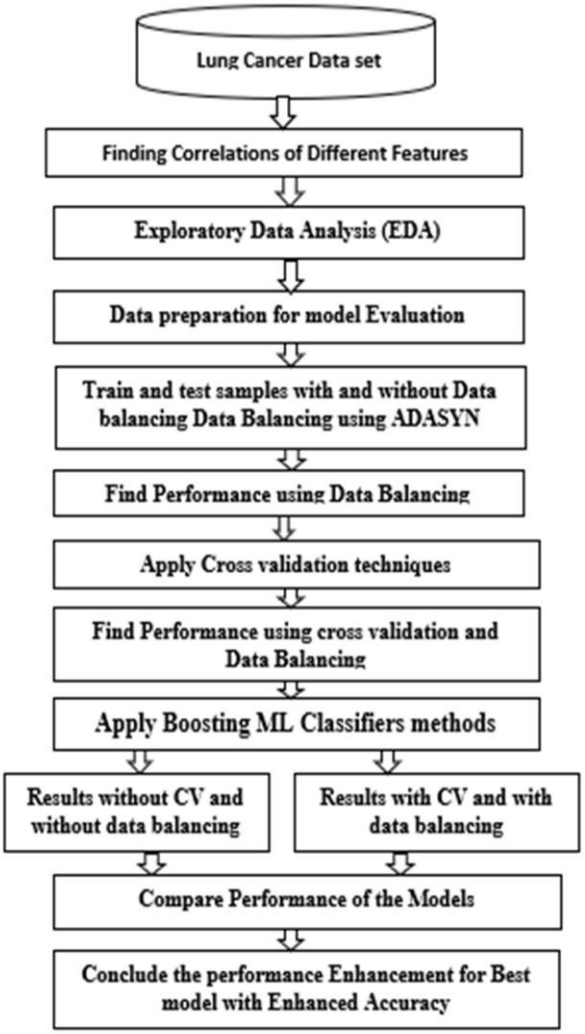

Fig. 2. System Block Diagram / Flow Diagram

5. RESULT AND DISCUSSION

In this work, we employed a wide range of categorization algorithms to make lung cancer prognoses. Five boosting machine learning classification models for lung cancer prediction A prediction model is being built using Python, a programming language, and the Scikit-learn module to detect lung cancer in people at an early stage. Based on a variety of factors, the code predicts the risk of lung cancer using five distinct machine learning algorithms: XGBoost, Adaboost, Catboost, LGBM, and gradient boosting classifier. The dataset used in the code includes various columns such as gender, age, smoking, yellow_fingers (YFN), anxiety (ANX), peer_pressure (PPR), chronic_disease (CDG), fatigue (FTG), allergy (ALG), wheezing (WHZ), alcohol_consuming (ALC), coughing (CHG), shortness_of_breath (SBG), swallowing difficulty (SWD), chest_pain (CHP), and lung_cancer. It is possible for the models for prediction to offer reliable estimates of a patient's risk of acquiring lung cancer. This is accomplished by analysing these data and using machine learning algorithms to recognize correlations and associations. It is necessary to carry out six different jobs in order to carry out the analysis of the outcomes. Importing libraries and datasets, finding correlations, and completing the first task.

Both the GENDER and LUNG_CANCER characteristics in this dataset are classified as object data types. Consequently, let's use the Label Encoder from sklearn to transform them into numerical values right now. This utility class, known as Label Encoder, is designed to assist in normalizing labels in such a way that they only include values that fall between 0 and n_classes-1. The transformation of non-numerical labels into numerical labels is another use of this tool, provided that the labels in question are hashable and comparable. In addition, let's set the value of every other property to YES=1 and NO=0. According to the correlation table or matrix, the degree of association between ANXIETY and YELLOW_FINGERS is more than fifty percent. Let us thus design a new feature that combines both of them.

Task 2 - Exploratory Data Analysis (EDA):

With regard to these characteristics, the value of zero may be seen as a missing value; hence, we will be replacing them with Nan and will make certain preparations in order to fill in these missing values. Now, we have some values that are NULL. We are going to use some kind of approach to fill in the missing data. Most of the characteristics have a gaussian distribution that is somewhat loose, which is beneficial for us.

Task - 3 - Data Preparation for model evaluation:

When it comes to transforming raw data into information that can be used, the first and most fundamental stage is data preparation. Raw data may, in general, have errors such as being incomplete, redundant, or noisy. In order to generate machine learning models, it is possible to fix all of these concerns that have been discussed via the process of data preparation. Checking the Outliers is something that is done in this Task. The duplicate values are dropped. NULL values are also corrected.

Task-4: Data Balancing:

Through the process of ADASYN as data balancing a dataset, it is possible to facilitate the training of a model by preventing the model from being biased towards a certain class. To put it another way, the model will no longer prefer the class that constitutes the majority simply since it includes more data. The different performance metrics measured are Precision (PR), Recall (RC), F1-Score (FS), Support (SP) for different such as Class-0 (CL-0) , Class-1 (CL-1) , Evaluation metrics like Accuracy (AC) , Macro Average (MA) , Weighted Average (WA). Table I shows the results for distinct parameters.

Table I. Analysis with Kfold CV -With and without Data Balance

	Precision	Recall	F1-Score	Accuracy	CVS	ROC-AUC
	Without Data Balancing					
GB	0.77	0.8	0.79	0.9	0.94	0.8
XGB	0.77	0.8	0.79	0.9	0.94	0.8
LGBM	0.43	0.5	0.47	0.87	0.81	0.5
CATBOOST	0.78	0.6	0.63	0.88	0.82	0.6
ADABOOST	0.77	0.8	0.79	0.9	0.94	0.8
	With Data Balancing					

GB	0.97	0.98	0.98	0.98	0.97	0.98
XGB	0.97	0.98	0.97	0.97	0.97	0.97
LGBM	0.87	0.88	0.87	0.88	0.88	0.87
CATBOOST	0.95	0.94	0.94	0.94	0.91	0.93
ADABOOST	0.97	0.97	0.97	0.97	0.97	0.97

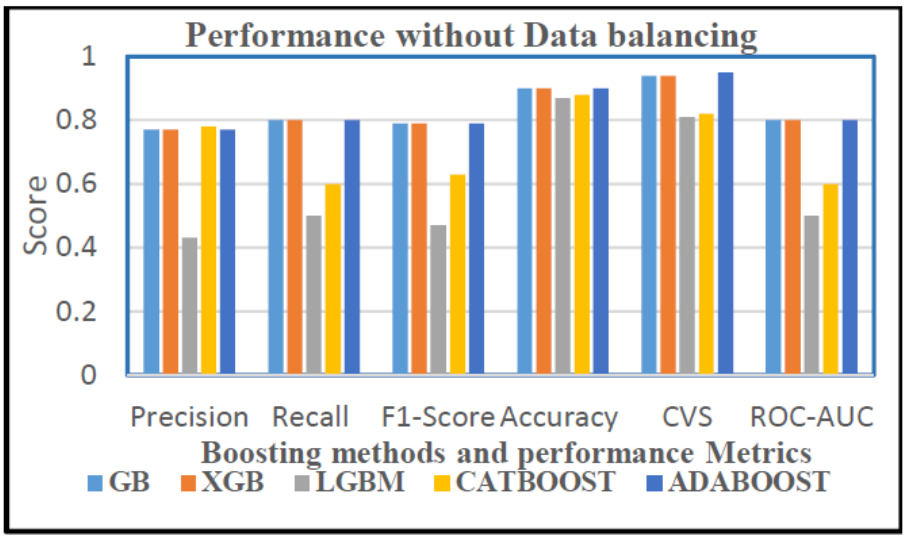

Fig. 3. Performance without data balancing

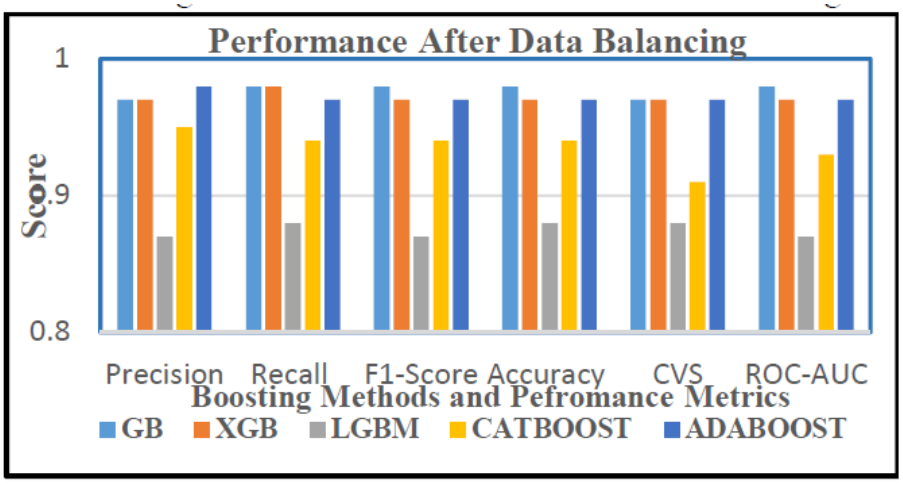

Fig. 4. Performance with data balancing

This shows that overall performance metrics parameters like Precision, Recall, F1-Score, Accuracy, Cross validation score and ROC-AUC Score is Improved after making use of data balancing as shown in figure 3 and 4. This further also shows that Gradient boosting method outperforms with respect to other boosting algorithms for all performance parameter.

6. CONCLUSION

The main objective of this proposed research work is to investigate the effectiveness of various Boosting strategy models by using a dataset including lung cancer as an input to identify a specific lung cancer illness prediction. To improving the accuracy of lung disease prediction, this study effort is being carried out to determine which boosting approaches have the highest level of performance and which cross validation approach is the most effective. Several different performance

indicators, including precision, recall, accuracy, and F1-score, are used to assess the performance evaluation of the overall technique.

To this study, we make use of the traditional Lung Cancer Dataset and implement several Boosting algorithms-based Machine Learning categorization approaches to it. These techniques include Gradient Boost (GB), XGBOOST (XGB), ADABOOST, CATBOOST (GB), and LightGBM (LGBM) along with K-Fold cross-validation methods and ADASYN as data balancing approach are all integrated with the Boosting algorithms to examine how the accuracy of lung cancer prediction is affected by hybrid combinations of Boosting approaches, K-Fold Cross validation, and data balancing methodology. This study has demonstrated a hybrid approach that can accurately predict lung cancer. According to the study's findings, a hybrid combination of one of the machine learning classifiers, the GB Model, which falls under the category of models based on boosting algorithms and makes use of ADASYN as a data balancing technique and K-Fold Cross validation, has remarkable accuracy and precision in predicting lung cancer.

REFERENCES

[1] Siegel, R.L.; Miller, K.D.; Jemal, A. “Cancer statistics”, 2020. CA Cancer J. Clin. 2020, 70, 7–30.

[2] Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. “Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 Countries”. CA Cancer J. Clin. 2021, 71, 209–249.

[3] Shah, R.; Sabanathan, S.; Richardson, J.; Mearns, A.; Goulden, C. “Results of surgical treatment of stage i and ii lung cancer “.J. Cardiovasc. Surg. 1996, 37, 169–172.

[4] Kuan, K.; Ravaut, M.; Manek, G.; Chen, H.; Lin, J.; Nazir, B.; Chen, C.; Howe, T.C.; Zeng, Z. “Deep learning for lung cancer detection: Tackling the kaggle data science bowl 2017 challenge”. arXiv 2017, arXiv:1705.09435. Diagnostics 2023, 13, 2617 23 of 27.

[5] I. D. Mienye, Y. Sun, and Z. Wang, “Improved Predictive Sparse Decomposition Method With Densenet For Prediction of Lung Cancer,” International Journal of Computing. Research Institute for Intelligent Computer Systems, pp. 533–541, Dec. 30, 2020. doi: 10.47839/ijc.19.4.1986.

[6] 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.

[7] J. D. Minna, J. A. Roth, and A. F. Gazdar, “Focus on lung cancer,” Cancer Cell, vol. 1, no. 1. Elsevier BV, pp. 49–52, Feb. 2002. doi: 10.1016/s1535-6108(02)00027-2.

[8] D. B. Snoke, G. S. Atwood, E. R. Bellefleur, A. M. Stokes, and M. J. Toth, “Body composition alterations in patients with lung cancer,” American Journal of Physiology-Cell Physiology, vol. 328, no. 3. American Physiological Society, pp. C872–C886, Mar. 01, 2025. doi: 10.1152/ajpcell.01048.2024.

[9] Ling S, Hu Z, Yang Z, Yang F, Li Y, Lin P, et al. “Extremely high genetic diversity in a single tumor point to prevalence of nondarwinian cell evolution”. Proc Natl Acad Sci U S A 2015;112: E6496–505.

[10] International Cancer Genome Consortium, Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, et al. “international network of cancer genome projects”. Nature 2010; 464:993–8.

[11] Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, et al. “The Cancer Genome Atlas Pan-Cancer analysis project”. Nat Genet 2013; 45:1113–20.

[12] D. B. Snoke, G. S. Atwood, E. R. Bellefleur, A. M. Stokes, and M. J. Toth, “Body composition alterations in patients with lung cancer,” American Journal of Physiology-Cell Physiology, vol. 328, no. 3. American Physiological Society, pp. C872–C886, Mar. 01, 2025. doi: 10.1152/ajpcell.01048.2024.

[13] Schabath MB, Cote ML. Cancer progress and priorities: lung cancer cancer. Epidemiol Biomarkers Prev. 2019;28(10):1563–79.

[14] Wang R, Dai W, Gong J, Huang M, Hu T, Li H, Lin K, Tan C, Hu H, Tong T, Cai G. “Development of a novel combined nomogram model integrating deep learning pathomics, radiomics and immune score to predict postoperative outcome of colorectal cancer lung metastasis patients”. Journal of Hematology & Oncology volume 15, Article number: 11 (2022)

[15] Mu Y, Zhou Y, Wang Y, Li W, Zhou L, Lu X, Gao P, Gao M, Zhao Y, Wang Q, Wang Y, Xu G. “Serum metabolomics study of non-smoking female patients with non-small cell lung cancer using gas chromatography-mass spectrometry”. J Proteome Res. 2019; 18:2175–84.

[16] Puneet and A. Chauhan, "Detection of Lung Cancer using Machine Learning Techniques Based on Routine

Blood Indices," 2020 IEEE International Conference for Innovation in Technology (INOCON), 2020, pp. 1-6, doi: 10.1109/INOCON50539.2020.9298407.

[17] M. I. Faisal, S. Bashir, Z. S. Khan and F. Hassan Khan, "An Evaluation of Machine Learning Classifiers and Ensembles for Early-Stage Prediction of Lung Cancer," 2018 3rd International Conference on Emerging Trends in Engineering, Sciences and Technology (ICEEST),2018, pp. 1-4, doi: 10.1109/ICEEST.2018.8643311.

[18] A. Safiyari and R. Javidan, "Predicting lung cancer survivability using ensemble learning methods," 2017 Intelligent Systems Conference (IntelliSys), 2017, pp. 684-688, doi: 10.1109/IntelliSys.2017.8324368 19. Mamun M., Farjana A., al Mamun M., Ahammed M.S. 2022 IEEE World AI IoT Congress, AIIoT 2022. 2022. "Lung cancer prediction model using ensemble learning techniques and a systematic review analysis"; pp. 187–193. [CrossRef]

[19] Patra R. "Prediction of lung cancer using machine learning classifier"; Communications in Computer and Information Science. Vol. 1235. CCIS; 2020pp. 132– 142.

[20] Jin-ah Sim J., et al. "The major effects of health-related quality of life on 5-year survival prediction among lung cancer survivors: applications of machine learning". Sci Rep. Dec. 2020;10(1) doi: 10.1038/s41598-020-67604-3.

[21] 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.

[22] 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.

[23] 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)

[24] Friedman, J. (2001). "Greedy boosting approximation: a gradient boosting machine". Ann. Stat. 29, 1189–1232. doi: 10.1214/aos/1013203451 https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

[25] Berrar, D. "Cross Validation"; Data Science Laboratory, Tokyo Institute of Technology: Tokyo, Japan, 2018

[26] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. "Scikit-learn: machine learning in Python". J Mach Learn Res. 2011; 12:2825– 30

[27] 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.

[28] H. B. M. Mohammed and N. Cavus, "Utilization of Detection of Non-Speech Sound for Sustainable Quality of Life for Deaf and Hearing-Impaired People: A Systematic Literature Review," Sustainability, vol. 16, no. 20. MDPI AG, p. 8976, Oct. 17, 2024. doi: 10.3390/su16208976.

[29] N. O. Beese et al., "Feel me, hear me: vibrotactile and auditory feedback cues in an invisible object search in virtual reality," Behaviour & Information Technology. Informa UK Limited, pp. 1–12, Feb. 13, 2025. doi: 10.1080/0144929x.2025.2459248.