An Extensive Analysis of Machine Learning Techniques for Predicting the Onset of Lung Cancer

D P Singh

Amity University Uttar Pradesh Greater Noida Campus

Abstract:

This research paper conducts an extensive analysis of various machine-learning techniques employed in predicting the onset of lung cancer. Lung cancer is a leading cause of mortality worldwide, emphasizing the critical need for accurate and timely prediction methods. Several machine-learning algorithms, such as SVM, RF, neural networks, and ensemble techniques, have been explored in this research. Through a comprehensive review and comparative study, this paper evaluates the effectiveness of different machine learning models in predicting lung cancer onset based on various input features, such as demographic data and medical history. The effectiveness of these methods is assessed using important metrics like accuracy, sensitivity, F1-score, computational efficiency, and the area under the receiver operating characteristic curve (AUC-ROC).Continuing technological progress is changing how healthcare operates, with the incorporation of machine learning techniques showing great promise in detecting issues early and predicting outcomes.

Furthermore, feature selection methods and data pre-processing techniques are explored to enhance prediction accuracy and reduce computational complexity. The results of this study offer important insights into the efficiency of machine learning for lung cancer prediction and suggest recommendations for future research.

Keywords:

Machine Learning, Lung Cancer Prediction, Early Detection, Comparative Study, Personalized Medicine, Feature Selection, Data Pre-Processing, Accuracy, Sensitivity, AUC-ROC.

1. Introduction

Lung cancer is one of the primary causes of cancer-related fatalities worldwide, representing a substantial share of cancer deaths. The high mortality rate is largely due to late diagnosis, as the symptoms of lung cancer often do not appear until the disease is in its advanced stages. Early detection is critical for improving survival rates, making the development of reliable predictive tools an essential goal in cancer research.

Lung cancer is the most lethal and hazardous form of cancer. Smoking is the basic risk factor for lung cancer [28, 29, 30, 31, 36], and it accounts for 85 out of 100 people dying every year [31]. Although people who do not smoke have a lower risk factor, the smoke of other smokers [30] may still affect them. Uranium is a metallic chemical element, which breaks down, with time, to form radon gas, which spreads in the air and water causing pollution and great harm to the lungs [31]. Lung cancer risk degree increases when there are cases of lung cancer in relatives, and this may be due to a common environment, genes or both [31]. In addition, the history of chronic pulmonary diseases is associated with lung cancer [31,36]. Prognostic models to predict cancer have been developed in many cases, including the incorporation of these tools for patient selection and pre-treatment stratification into clinical trials [10]; some of these tools predicted lung cancer [4].

Lung cancer is among the most commonly diagnosed types of cancer and is the primary cause of cancer-related deaths globally. About 2.20 million new patients are diagnosed with lung cancer each year [46], and 75% of them die within five years of diagnosis [42].High intra-tumor heterogeneity (ITH) and complexity of cancer cells giving rise to drug resistance make cancer treatment more challenging [17]. Over the past decades, the continuous evolution of technologies in cancer research has contributed to many large collaborative cancer projects, which have generated numerous clinical, medical imaging, and sequencing databases [7, 11, 12]. These databases facilitate researchers in investigating comprehensive patterns of lung cancer from diagnosis, treatment, and responses to clinical outcomes [18].

Recent research in omics analysis, including genomics, transcriptomics, proteomics, and metabolomics, has enhanced our research tools and capabilities. Cancer research is transitioning towards incorporating various data types and large-scale datasets. However, working with varied and high-dimensional data types for clinical tasks demands considerable time and expertise, even when utilizing dimension reduction techniques like matrix and tensor factorizations [6, 8, 24, 37,]. Additionally, analyzing the rapidly expanding cancer-related databases presents a significant challenge for researchers. As a result, leveraging machine learning (ML) models to automatically understand the inherent traits of various data types to aid physicians in their decision-making has gained significance. ML is a subgroup of artificial intelligence (AI) that focuses on making predictions by identifying patterns in data using mathematical algorithms [47]. It has served as an assisting tool in cancer phenotyping and therapy for decades [1,5,41,43], and has been widely implemented in advanced approaches for early detection, cancer type classification, signature extraction, tumor microenvironment (TME) deconvolution, prognosis prediction, and drug response evaluation [14,38, 44,45, 48, 49,50,53].

Lung cancer is the leading cause of cancer-related deaths, accounting for approximately 1.8 million deaths in 2020 [54,55]. The cancer survival period is the timeframe from when a person is diagnosed with cancer until their death due to the disease [51]. Additionally, survival analysis is beneficial for clinicians, researchers, patients, and policymakers. Developing an accurate and robust model is crucial for identifying lung cancer survival rates [56]. Various ML algorithms have been developed for clinical applications, including random forests (RFs), ensemble algorithms, Naive Bayesian (NB) classifiers, Support vector machines (SVMs), neural networks (NNs), Decision Trees (DTs), and a number of proprietary algorithms [52]. Machine learning (ML) techniques can connect different clinical characteristics of cancer patients to their survival rates. Moreover, ML helps alleviate the workload of healthcare professionals and minimizes the potential for human error. The impressive effectiveness of ML has made it an appealing and inspiring resource for those in the healthcare field. These techniques enable the creation of predictive models utilizing cancer data to forecast survival outcomes. Despite this, no current technique qualifies for application to a specific dataset [57].

Machine learning (ML) has emerged as a powerful tool in the healthcare domain, offering advanced methods for analyzing large and complex datasets. In the context of lung cancer, ML techniques can be employed to identify patterns and features that are predictive of the disease's onset, potentially leading to earlier diagnosis and better patient outcomes.

This paper presents an extensive analysis of machine learning techniques applied to the prediction of lung cancer onset. We explore a variety of ML approaches, including supervised learning methods such as logistic regression, support vector machines, and neural networks; unsupervised methods like clustering and principal component analysis; and ensemble methods such as random forests and gradient boosting machines. Our goal is to assess the effectiveness of these techniques in predicting lung cancer and to identify the most promising approaches for clinical application.

2. Literature review

Lung cancer has become a significant focus for researchers in both oncology and the field of AI-driven medical assistance. Some studies have designed systems for the detection and diagnosis of lung cancer [21,25], while others have focused on early lung cancer diagnosis [13, 15, 16, 26, 33]. Studies about the diagnosis of lung cancer have been based on techniques such as fuzzy logic8 and neural networks [39]. Other studies have used hybrid neuro-fuzzy techniques [26,34]. However, these methods cannot effectively create a reliable medical diagnostic system as the size of databases continues to grow, rendering them less dependable. There are studies based on advanced machine learning concepts, such as decision trees [16,26,34,40], which have demonstrated higher reliability compared to those old systems. Hanai and others introduced prognostic models for Non-Small-Cell Lung Cancer (NSCLC) based on neural networks [2]. Kattan and Bach introduced a study on the variations in lung cancer risk among smokers based on many factors [3]. Ramachandran and others built an early prevention system for lung cancer based on data mining in which they used 11 different factors [15]. They carried out experiments with a database of 746 samples, but did not provide any information about the source of this database. In 2014, Thangaraju and others also used data mining techniques to predict the risk factor of lung cancer [16]. They employed Bayes Trees and Decision Tables for both clustering and classification, conducting experiments with 303 samples. Manikandan and others designed a hybrid neuro-fuzzy system for the prediction of lung cancer based on 11 symptoms [33]. Arulananth and Bharathi defined the symptoms that can be used for lung cancer prediction [35]. They made a distinction between diagnostic factors and the symptoms that signal the presence of cancer. The diagnostic symptoms were categorized based on age, sex, family history of cancer, smoking habits, radiation exposure, radon exposure, chemical exposure, and air pollution. In contrast, the symptoms indicating the presence of cancer included chronic cough, hemoptysis, chest pain, weight loss, fatigue, chronic lung inflammation, wheezing, difficulty swallowing, and anorexia.

In 2018, Senthil and Ayshwaya used neural networks and evolutionary algorithms to define the risk degree of lung cancer based on risk factors [39]. Recently, in 2018, Markaki and others built a clinical risk prediction model for lung cancer based on smoking symptoms [40]. Some other studies used advanced machine learning algorithms, such as random trees and random forests, which were very useful for the classification of big databases [27]. On the other hand, others have relied on radiotherapy image processing techniques to determine whether lung cancer is present or not [22]. Other researches focused on the prediction of the mortality of people with NSCLC in the U.S. Military Health System [9]. Cassidy concluded that for building a good lung cancer risk prediction model, it was preferable to seek other factors in addition to smoking and age [4]. This study develops a lung cancer prediction tool that incorporates various risk factors and their details. It also examines the symptoms and their impact on lung cancer. To create a robust international prediction tool, the research takes into account both local and global studies and reports.

3.Machine Learning Models: Several algorithms were utilized in constructing the prediction system, categorized as[58,59,60]:

3.1.Logistic Regression: Logistic regression is a statistical method mainly utilized for binary classification, aiming to estimate the likelihood of a binary result (e.g., yes/no, 0/1, true/false). The logistic regression model can be mathematically expressed as follows [58, 59, 60]:

The probability is modeled using the logistic (sigmoid) function.

Probability of Class 1: $P\left(y = \frac{1}{x} = \sigma(z)\right) = \frac{1}{1+e^{-z}}$

Where: $\sigma(z)$ is the sigmoid function.

z is the linear combination of input features and model parameters:$z = w^T + b$.

$w = [w_1, w_2, w_3, \ldots . w_n]$ are the weights (coefficients) associated with each feature.

b is the bias (intercept) term.

$x = [x_1, x_2, x_3, \ldots \ldots x_n]$ are the input features.

Probability of Class 0: $P\left(y = \frac{0}{x}\right) = 1 - P\left(y = \frac{1}{x}\right) = 1 - \sigma(z)$

Decision Boundary: The model classifies the input x as class 1 if $\left(y = \frac{1}{x}\right) \geq 0.5$, and as class 0 otherwise. The decision boundary occurs where the probability is exactly 0.5,which corresponds to z=0: $w^T + b = 0$

Cost Function: To train the logistic regression model, we minimize a cost function. The log-loss, also known as binary cross-entropy loss, is the most frequently utilized cost function in logistic regression:

$$J(w,b) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^i log\left(P(y^i = 1/x^i)\right) + (1-y^i) log\left(1 - P(y^i = 1/x^i)\right)\right]$$

Where m represents the total number of training examples. y^i denotes the true label for the i-th example. $P(y^i = 1/x^i)$ indicates the predicted probability that the i-th example belongs to class 1.

Gradient Descent Optimization: The model parameters w and b are updated iteratively using gradient descent to minimize the cost function: $w \coloneqq w - \alpha \nabla_w J(w,b)$

$$b \coloneqq b - \alpha \nabla_b J(w,b)$$

Where: α is the learning rate.

$\nabla_w J(w,b)$ and $\nabla_b J(w,b)$ are the gradients of the cost function with respect to the weights and bias, respectively.

Prediction: Finally, once the model is trained, predictions are made by computing the probability P(y=1|x) for a new input x and assigning the class label based on the decision boundary:

$$\breve{y} = \begin{cases} 1 & if\ \sigma(z) \geq 0.5 \\ 0 & if\ \sigma(z) < 0.5 \end{cases}$$

This is the basic mathematical framework behind logistic regression.

3.2.Gaussian Naive Bayes (GNB): Gaussian Naive Bayes assumes that the features follow a Gaussian (normal) distribution.

Model Assumption: Features are conditionally independent given the class label.

Probability Model:

$$P(C|X_1, X_2, X_3 \dots \dots \dots X_n) = \frac{P(C)\prod_{i=1}^{n} P(X_i|C)}{P(X_1, X_2, X_3 \dots \dots \dots X_n)}$$

Where C is the class, X_1,X_2,…,Xn are the feature values.

Gaussian Assumption: The conditional probability for feature X_i given class C is modeled as: $P(X_i|C) = \frac{1}{\sqrt{2\pi\sigma_C^2}} e^{\left(-\frac{(X_i-\mu_C)^2}{2\,\sigma_C^2}\right)}$

Where μ_C and $\sigma_C$are the mean and standard deviation of the feature for class C

3.3. Bernoulli Naive Bayes (BNB): Bernoulli Naive Bayes is used for binary/Boolean features (0/1 values).

Model Assumption: Like GNB, Bernoulli Naive Bayes assumes conditional independence between features.

Probability Model: $P(C|X_1, X_2, X_3 \dots \dots \dots X_n) = \frac{P(C)\prod_{i=1}^{n} P(X_i|C)}{P(X_1, X_2, X_3 \dots\dots\dots X_n)}$

$$P(X_1|C) = P_C^{X_i}(1-p_c)^{(1-X_i)}$$

where p_c is the probability that feature X_i =1 in class C.

3.4.Support Vector Machine (SVM): SVM is a type of supervised learning model that aims to identify the optimal hyperplane to divide data into distinct classes effectively.

Objective: Identify a hyperplane that maximizes the separation margin between two classes. Model: The decision boundary is defined by: $w^T x + b = 0$

where w represents the weight vector, x stands for the input vector, and b denotes the bias term.

Optimization: The objective is to increase the separation between the two classes, subject to the constraint that all points are classified correctly: $\min_{w,b} \frac{1}{2} \|w\|^2$. Subject to $y_i(w^T x + b) \geq 1$ for all i , where $y_i\{-1,1\}$is the class label.

3.5. Decision Trees: A decision tree is a highly effective tool in supervised learning, used for both classification and regression tasks. It is organized like a tree diagram, where each internal node corresponds to a test on a feature, each branch represents a possible result of that test, and each leaf node signifies a class label. The tree is constructed by repeatedly dividing the training data into smaller subsets based on attribute values, halting when certain conditions are met, such as reaching a maximum tree depth or a minimum number of samples needed to split a node. Mathematically, it's structured as:

Entropy (used in classification trees): Entropy measures the impurity in a split. For a node with binary classification (0 or 1), entropy E is defined as:

$$E(S) = -p_1 log_2(p_1) - p_0 log_2(p_0)$$

Where $p_1$the proportion of is class 1 in set S and p_0 is the proportion of class 0.

Gini Index (another impurity measure): $Gini(S) = 1 - p_1^2 - p_0^2$

Recursive Splitting: The tree recursively splits data at each node by selecting a feature and threshold that minimizes impurity (e.g., entropy or Gini) at the child nodes.

Cost Function: For a regression tree, the cost function to minimize is usually the mean squared error (MSE): $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_1)^2$,where y_i is the true label, and $\hat{y}_1$ is the predicted label for i-th the sample.

3.6.Random Forest: Random Forest, a widely-used ensemble learning method involving decision trees, creates a 'forest' of multiple trees. These trees are usually trained with the 'bagging' technique, which merges multiple models to improve the overall result. Random Forest boosts the performance of Decision Trees by reducing variance, achieved by growing more trees and introducing more randomness into the model. Rather than always choosing the most significant feature for splitting nodes, it selects the best feature from a random subset of features, leading to a more robust model.

Random Forest Regression is a machine learning ensemble method capable of managing both regression and classification tasks by utilizing multiple decision trees and applying Bootstrap and Aggregation, commonly known as bagging. Rather than relying on a single decision tree, this approach combines the outputs of several trees to produce the final result. In Random Forest, numerous decision trees act as the core learning models.

3.7.Gradient Boosting:

Gradient Boosting is an iterative process where weak learners (usually decision trees) are added sequentially to minimize a loss function. Its mathematics is based on functional gradient descent.

Loss Function: Let $L(y, \hat{y})$be the loss function to be minimized (e.g., log loss for classification, MSE for regression). The goal is to find a function F(x) such that the predictions F(x) minimize this loss: $F(x) = arg \min_{F(x)} \sum_{i=1}^{n} L(y_i, F(x_i))$

Additive Model: Gradient Boosting builds an additive model: $F_m(x) = F_{m-1}(x) + \alpha . h_m(x)$

where $F_{m-1}(x)$ is the prediction from the previous iteration, $h_m(x)$ is the new decision tree (or weak learner), and α is the learning rate.

Gradient Descent: The new model $h_m(x)$ is trained to match the negative gradient of the loss function relative to the current predictions: $h_m(x) = arg \min_{F(x)} \sum_{i=1}^{n} \left[-\frac{\partial L(y_i,F_{m-1}(x))}{\partial F_{m-1}(x)} - h(x_i)\right]^2$

Final Prediction: After M iterations, the final prediction is: $\hat{y} = F_m(x) = \sum_{m=1}^{M} \alpha h_m(x)$.

3.8.K-Nearest Neighbours (KNN): KNN is an instance-based learning algorithm that is non-parametric and can be applied to both classification and regression tasks.

To predict the class or value of a new data point, KNN measures the distance between the new point and all points in the training dataset.

The Euclidean Distance is the most frequently used distance metric:

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

where x and y are two points in an n-dimensional space.

Prediction: For classification, the algorithm finds the k-nearest neighbours to the new data point and assigns the class based on a majority vote of the neighbours' labels.

For regression, it predicts the value by taking the mean (or weighted average) of the nearest neighbours' values.

3.9.Extreme Gradient Boosting (XGBoost): XGBoost is a gradient boosted decision tree framework optimized for efficiency and speed. It employs a collection of decision trees, with each new tree aiming to rectify the mistakes made by the previous ones by placing greater emphasis on samples that were misclassified.

Gradient Descent: XGBoost minimizes the loss function by gradient descent. For a loss function $L(y,\hat{y})$, where y is the true label and $\hat{y}$ is the prediction, the next tree tries to minimize: $L = \sum_{i=1}^{n} Loss(\ y,\hat{y}) + \sum_{k=1}^{K} \Omega(f_k)$

where:

$\Omega(f_k)$ is a regularization term to prevent overfitting.

f_k is a weak learner, typically a decision tree.

The loss function is often chosen as logistic loss for classification and mean squared error for regression.

Regularization: To avoid overfitting, XGBoost uses both L_1 (Lasso) and L_2 (Ridge) regularization: $\Omega(f) = \Upsilon T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2$

where T represents the number of leaves, ω_j denotes the weight of leaf j and Υ λ are the regularization parameters.

3.10.Extra Tree Classifier (ETC): Extra Trees, also known as Extremely Randomized Trees, is an ensemble learning method akin to Random Forest, but it introduces greater randomness during the construction of the trees. Randomized Tree Splits: In the process of building the decision tree, Extra Trees randomly chooses a subset of features and thresholds for splitting rather than looking for the optimal split at each node. This randomness reduces the correlation between trees and can lead to better generalization.

Prediction: For classification, the final prediction is made by aggregating the votes from all trees (majority voting).For regression; the prediction is the average of the predictions from all trees.

Ensemble Model: $\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_t(x)$,

where T stands for the number of trees, and $h_t(x)$ is the prediction from the t-th tree for input x.

3.11.AdaBoost (Adaptive Boosting): AdaBoost is an ensemble technique that merges several weak classifiers to form a robust classifier. Weak classifiers are typically decision trees that consist of only a single split, also known as decision stumps. AdaBoost works by assigning more weight to misclassified instances at each iteration, so subsequent classifiers focus more on difficult cases.

Objective: Minimize the classification error by combining weak learners.

Steps:

1. Initialize weights $w_i = \frac{1}{N}$, where N is the number of samples.
2. For each classifier t:

Train a weak learner $h_t(x)$ with the weighted dataset.

Compute error $\varepsilon_t = \frac{\sum \omega_i \; I(y_i \neq h_t(x_i))}{\sum \omega_i}$ where I(·) is the indicator function. Compute classifier weight $\alpha_t = \frac{1}{2} \ln \left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$.

Update weights for misclassified points: $w_i \leftarrow w_i . e^{\alpha_t I(y_i \neq h_t(x_i))}$, and normalize the weights.

Final classifier is a weighted vote of weak learners: $H(x) = sign(\sum_t \alpha_t h_t(x))$

3.12. CatBoost (Categorical Boosting): CatBoost is a gradient boosting algorithm developed to handle categorical features natively. It is based on boosting decision trees and is highly efficient in dealing with high-cardinality categorical data. CatBoost reduces prediction shifts and applies an ordered boosting strategy.

Objective: Minimize a loss function (usually log-loss for classification or RMSE for regression) by iteratively building an ensemble of trees. During each iteration, train a decision tree to estimate the gradient of the loss function concerning the predictions from the existing ensemble of trees.

Update the model by adding the newly fitted tree to the ensemble:

$$F_m(x) = F_{m-1}(x) + \eta . h_m(x)$$

where $F_m(x)$ is the model after m-th iteration, $h_m(x)$ is the decision tree at iteration , and η is the learning rate.

CatBoost applies a permutation-driven strategy to avoid overfitting on categorical data by reducing the effect of target leakage. The ordered boosting formula can be written as:

$$F_m(x) = F_{m-1}(x) - \eta . \frac{\partial L}{\partial F_{m-1}(x)}$$

where L is the loss function and η is the learning rate.

3.13.Multi-layer Perceptron (MLP): MLP is a type of feedforward artificial neural network (ANN). It is made up of several layers of nodes, which include an input layer, hidden layers, and an output layer. Each node (or neuron) in one layer is linked to the neurons in the subsequent layer via weights.

Objective: Learn a mapping from input x to output y by minimizing a loss function (such as mean squared error or cross-entropy loss).

Mathematical formulation:

Forward pass: For each layer l, the output is: $z^{(l)} = \boldsymbol{W}^{(l)} \boldsymbol{a}^{(l-1)} + \boldsymbol{b}^{(l)}$

where $\boldsymbol{W}^{(l)}$ are the weights of layer l, $\boldsymbol{b}^{(l)}$ are the biases, and $\boldsymbol{a}^{(l-1)}$ is the activation of the previous layer.

Apply an activation function (·) (e.g., ReLU, Sigmoid) to get the output: $\boldsymbol{a}^{(l)} = g(z^{(l)})$

Backpropagation:

Compute the error at the output layer and propagate it backward to update the weights using gradient descent.

For each layer l, update the weights and biases: $\boldsymbol{W}^{(l)} \leftarrow \boldsymbol{W}^{(l)} - \eta . \frac{\partial L}{\partial \boldsymbol{W}^{(l)}}$

where η is the learning rate and L is the loss function.

Output: The output layer performs a linear combination of the activations of the previous layer, often applying a softmax function for classification tasks: $\hat{y}_i = \frac{e^{z_i}}{\sum_j e^{z_j}}$

where z_i are the raw outputs of the final layer.

4. Confusion Matrix in Machine Learning: A confusion matrix summarizes a machine learning model's performance on a test dataset, visually displaying both accurate and inaccurate predictions. It is commonly used to evaluate classification models, which assign categorical labels to input data.

This matrix is crucial for evaluating the performance of a classification model, as it offers detailed counts of true positives, true negatives, false positives, and false negatives. It allows for a more comprehensive analysis of the model's recall, accuracy, precision, and its overall capability to differentiate between classes by displaying the frequency of predicted outcomes in the test dataset[58,59,60].

4.1 Accuracy: Accuracy measures a model's effectiveness by calculating the ratio of correctly classified instances to the total number of instances.

$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$,

where TP= True positives, TN= True negatives, FP= False positives and FN= False negatives.

4.2 Precision: Precision refers to the accuracy of a model's positive predictions. It is measured by the ratio of true positive predictions to the total number of positive predictions made by the model.

$$Precision = \frac{TP}{TP + FP}$$

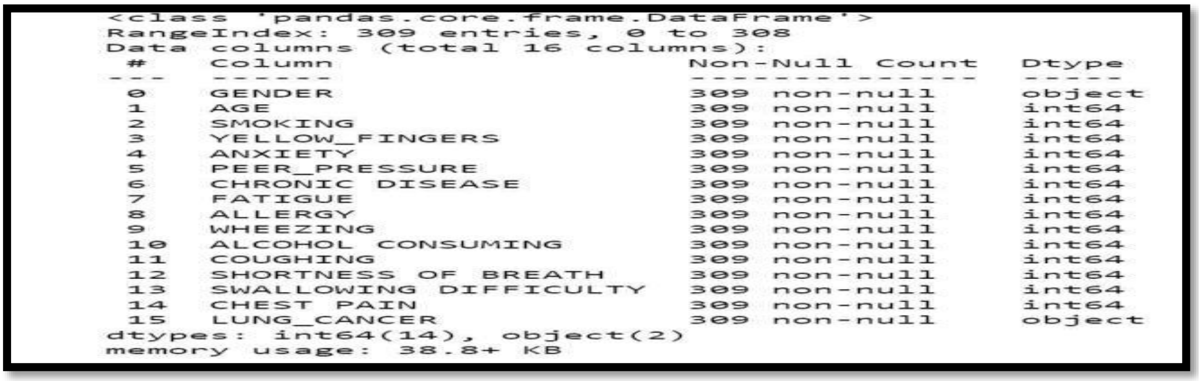
4.3 Recall: Recall measures how well a classification model can identify all the relevant instances within a dataset. It is calculated by dividing the number of true positive (TP) cases by the total number of true positives and false negatives (FN).

$$Recall = \frac{TP}{TP + FN}$$

4.4 Specificity: Specificity, an essential metric for evaluating classification models, particularly in binary cases, measures how accurately a model identifies negative instances, also known as the True Negative Rate.

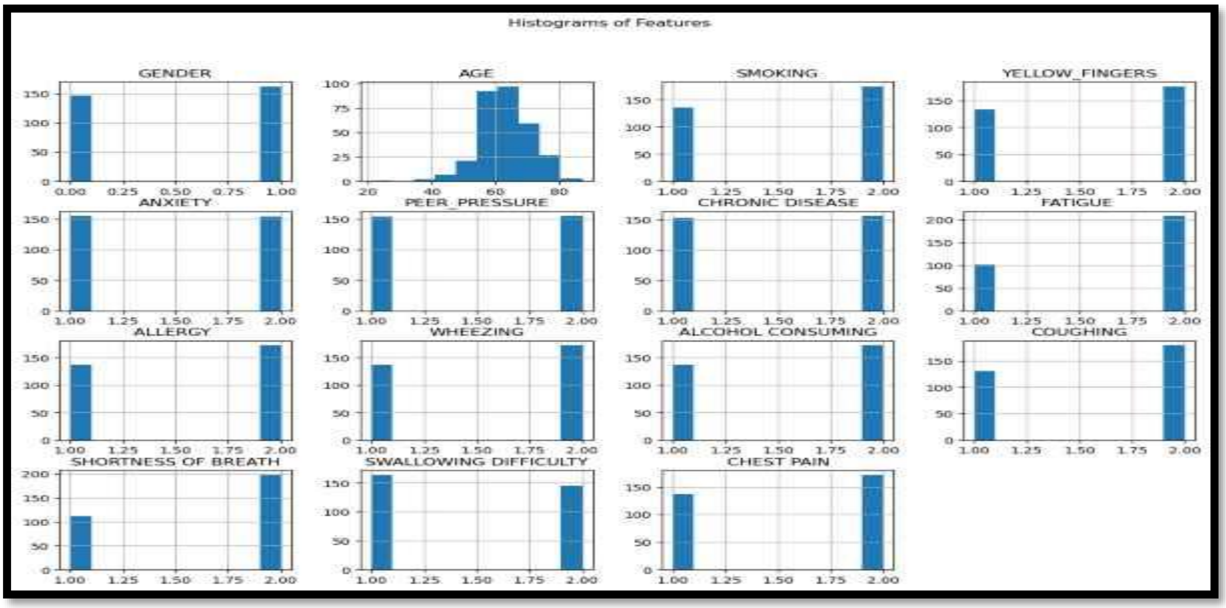
$$Specificity = \frac{TN}{TP + FP}$$

5. Data Cleaning and Feature Engineering: To assess lung cancer, we gathered data from (https://www.kaggle.com/datasets/akashnath29/lung-cancer-dataset?select=dataset.csv), which includes 15 clinical features, and used these to build predictive models for lung cancer detection. Table 1 indicates that the data contains no missing values.

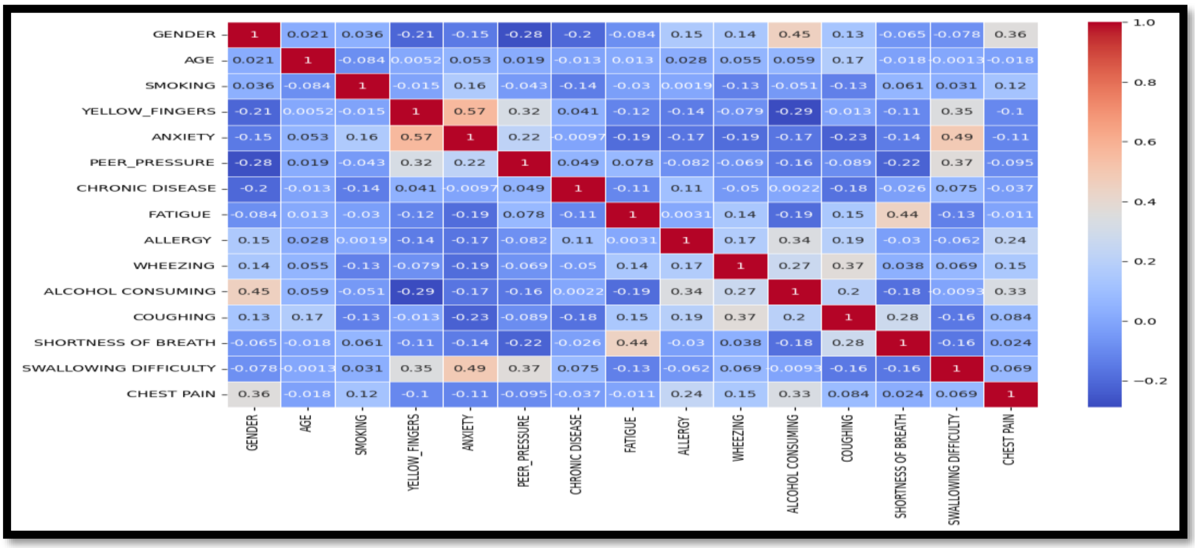


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 309 entries, 0 to 308
Data columns (total 16 columns):
 #   Column                 Non-Null Count  Dtype 
---  ------                 --------------  ----- 
 0   GENDER                 309 non-null    object
 1   AGE                    309 non-null    int64 
 2   SMOKING                309 non-null    int64 
 3   YELLOW_FINGERS         309 non-null    int64 
 4   ANXIETY                309 non-null    int64 
 5   PEER_PRESSURE          309 non-null    int64 
 6   CHRONIC DISEASE        309 non-null    int64 
 7   FATIGUE                309 non-null    int64 
 8   ALLERGY                309 non-null    int64 
 9   WHEEZING               309 non-null    int64 
 10  ALCOHOL CONSUMING      309 non-null    int64 
 11  COUGHING               309 non-null    int64 
 12  SHORTNESS OF BREATH    309 non-null    int64 
 13  SWALLOWING DIFFICULTY  309 non-null    int64 
 14  CHEST PAIN             309 non-null    int64 
 15  LUNG_CANCER            309 non-null    object
dtypes: int64(14), object(2)
memory usage: 38.8+ KB
```

In the histogram, we are illustrating the number of lung cancer cases associated with the clinical features:



The heatmap has been used to identify the relationships between all clinical features:

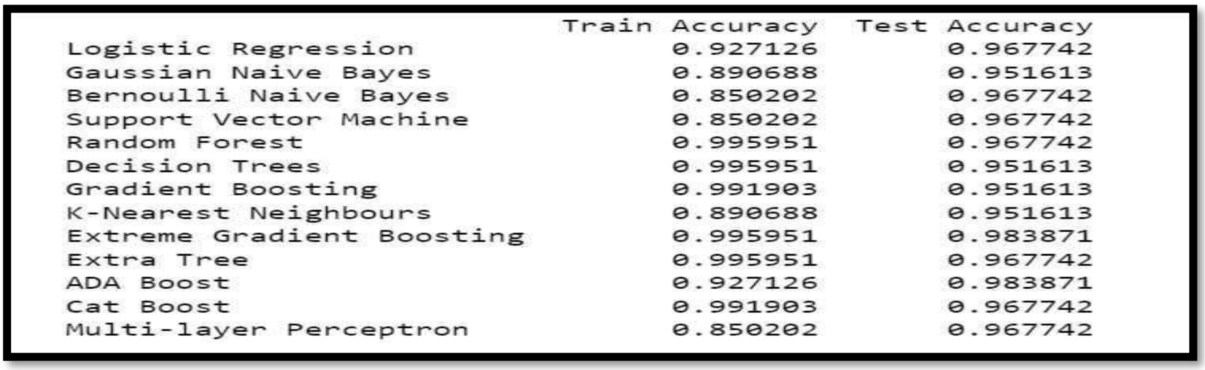


	GENDER	AGE	SMOKING	YELLOW_FINGERS	ANXIETY	PEER_PRESSURE	CHRONIC DISEASE	FATIGUE	ALLERGY	WHEEZING	ALCOHOL CONSUMING	COUGHING	SHORTNESS OF BREATH	SWALLOWING DIFFICULTY	CHEST PAIN
GENDER	1	0.021	0.036	-0.21	-0.15	-0.28	-0.2	-0.084	0.15	0.14	0.45	0.13	-0.065	-0.078	0.36
AGE	0.021	1	-0.084	0.0052	0.053	0.019	-0.013	0.013	0.028	0.055	0.059	0.17	-0.018	-0.0013	-0.018
SMOKING	0.036	-0.084	1	-0.015	0.16	-0.043	-0.14	-0.03	0.0019	-0.13	-0.051	-0.13	0.061	0.031	0.12
YELLOW_FINGERS	-0.21	0.0052	-0.015	1	0.57	0.32	0.041	-0.12	-0.14	-0.079	-0.29	-0.013	-0.11	0.35	-0.1
ANXIETY	-0.15	0.053	0.16	0.57	1	0.22	-0.0097	-0.19	-0.17	-0.19	-0.17	-0.23	-0.14	0.49	-0.11
PEER_PRESSURE	-0.28	0.019	-0.043	0.32	0.22	1	0.049	0.078	-0.082	-0.069	-0.16	-0.089	-0.22	0.37	-0.095
CHRONIC DISEASE	-0.2	-0.013	-0.14	0.041	-0.0097	0.049	1	-0.11	0.11	-0.05	0.0022	-0.18	-0.026	0.075	-0.037
FATIGUE	-0.084	0.013	-0.03	-0.12	-0.19	0.078	-0.11	1	0.0031	0.14	-0.19	0.15	0.44	-0.13	-0.011
ALLERGY	0.15	0.028	0.0019	-0.14	-0.17	-0.082	0.11	0.0031	1	0.17	0.34	0.19	-0.03	-0.062	0.24
WHEEZING	0.14	0.055	-0.13	-0.079	-0.19	-0.069	-0.05	0.14	0.17	1	0.27	0.37	0.038	0.069	0.15
ALCOHOL CONSUMING	0.45	0.059	-0.051	-0.29	-0.17	-0.16	0.0022	-0.19	0.34	0.27	1	0.2	-0.18	-0.0093	0.33
COUGHING	0.13	0.17	-0.13	-0.013	-0.23	-0.089	-0.18	0.15	0.19	0.37	0.2	1	0.28	-0.16	0.084
SHORTNESS OF BREATH	-0.065	-0.018	0.061	-0.11	-0.14	-0.22	-0.026	0.44	-0.03	0.038	-0.18	0.28	1	-0.16	0.024
SWALLOWING DIFFICULTY	-0.078	-0.0013	0.031	0.35	0.49	0.37	0.075	-0.13	-0.062	0.069	-0.0093	-0.16	-0.16	1	0.069
CHEST PAIN	0.36	-0.018	0.12	-0.1	-0.11	-0.095	-0.037	-0.011	0.24	0.15	0.33	0.084	0.024	0.069	1

6. Prediction of Lung Cancer: We examined the effectiveness of thirteen machine-learning algorithms. Afterwards, we evaluated how well these thirteen machine-learning models could serve as clinical decision support systems in forecasting lung cancer.

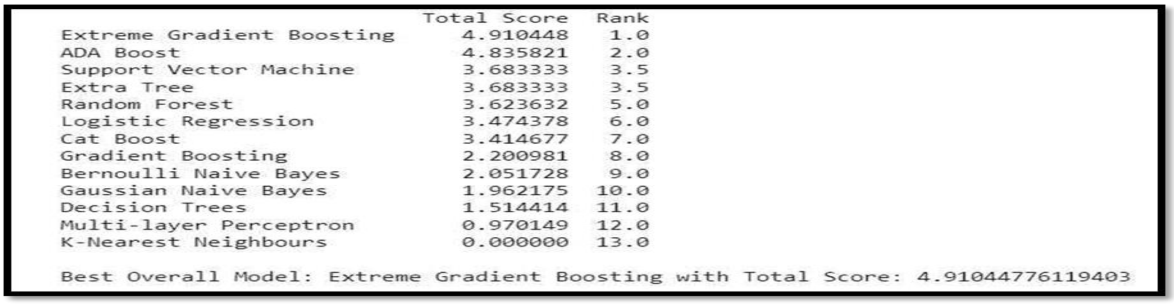
6.1. Examine the performance of Machine Learning models:

Partition the dataset into training, validation, and test sets to assess the model's performance. Standardize the data to ensure consistency, which is crucial for many machine learning algorithms. During model development, we randomly chose clinical features from 80% of the patients for training. A 3-fold internal cross-validation was then performed to evaluate the model’s predictive ability using this training data. Additionally, we tested the model's predictive accuracy on an independent sample for external validation, as shown in Table 2.



```
                              Train Accuracy  Test Accuracy
Logistic Regression                 0.927126       0.967742
Gaussian Naive Bayes                0.890688       0.951613
Bernoulli Naive Bayes               0.850202       0.967742
Support Vector Machine              0.850202       0.967742
Random Forest                       0.995951       0.967742
Decision Trees                      0.995951       0.951613
Gradient Boosting                   0.991903       0.951613
K-Nearest Neighbours                0.890688       0.951613
Extreme Gradient Boosting           0.995951       0.983871
Extra Tree                          0.995951       0.967742
ADA Boost                           0.927126       0.983871
Cat Boost                           0.991903       0.967742
Multi-layer Perceptron              0.850202       0.967742
```

6.2. Rank of Models for prediction Lung Cancer: Thirteen machine learning models were evaluated based on their predictive performance, and the most rank one was selected. Among these, the Extreme Gradient Boosting model rank is the best. The rankings of all the models are listed below:

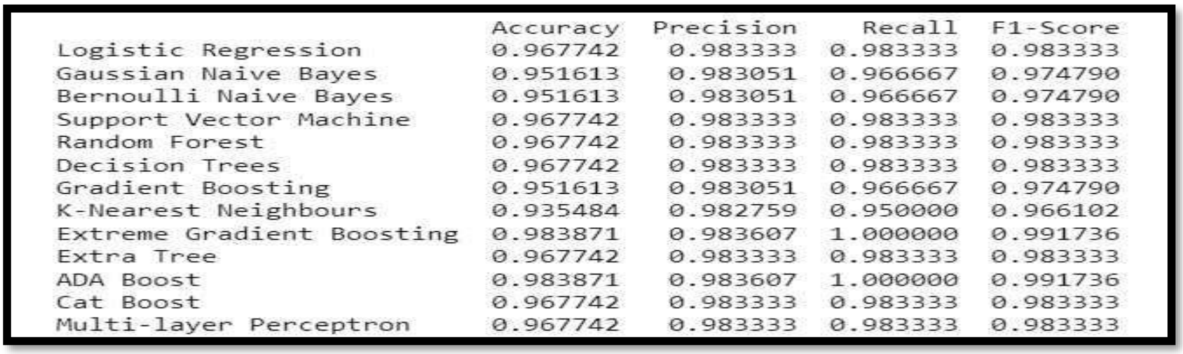


```
                           Total Score  Rank
Extreme Gradient Boosting     4.910448   1.0
ADA Boost                     4.835821   2.0
Support Vector Machine        3.683333   3.5
Extra Tree                    3.683333   3.5
Random Forest                 3.623632   5.0
Logistic Regression           3.474378   6.0
Cat Boost                     3.414677   7.0
Gradient Boosting             2.200981   8.0
Bernoulli Naive Bayes         2.051728   9.0
Gaussian Naive Bayes          1.962175  10.0
Decision Trees                1.514414  11.0
Multi-layer Perceptron        0.970149  12.0
K-Nearest Neighbours          0.000000  13.0

Best Overall Model: Extreme Gradient Boosting with Total Score: 4.91044776119403
```

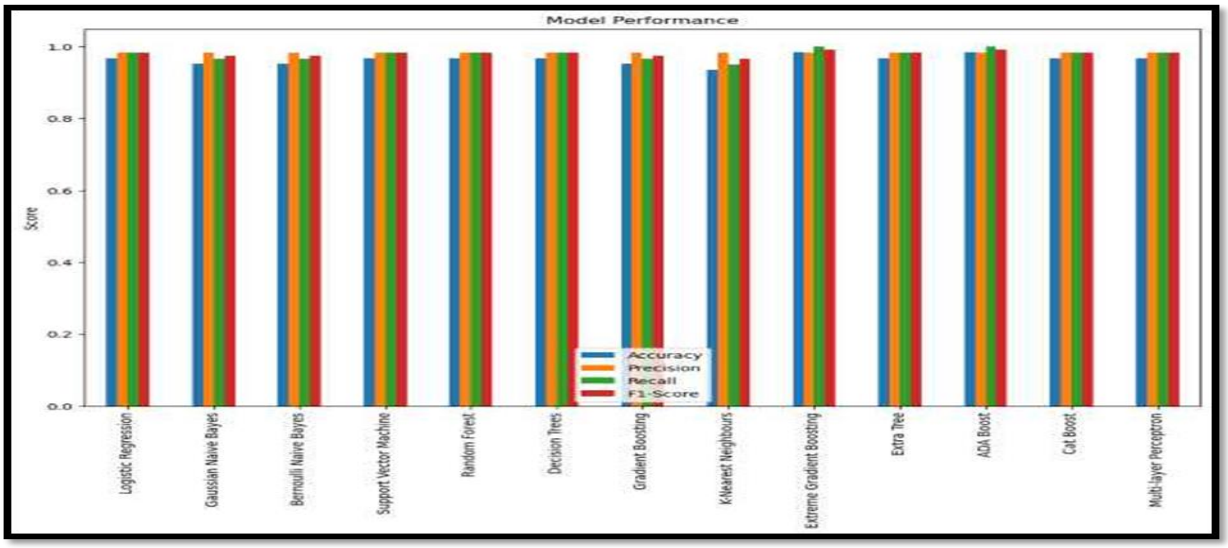
6.3. Choosing Final Model for Lung Cancer Prediction:

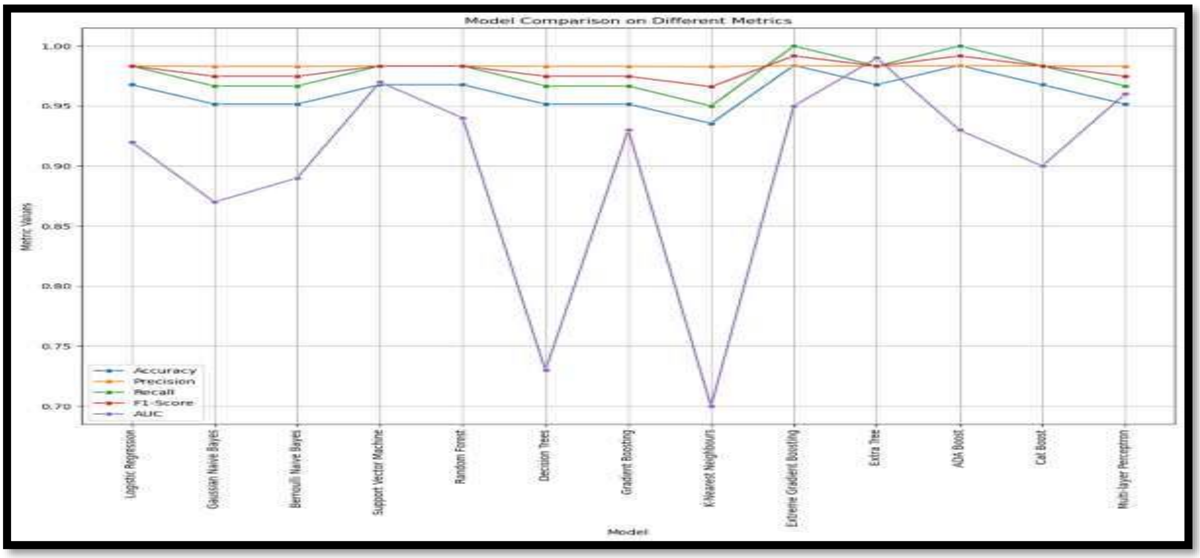
The Extreme Gradient Boosting model proved to be the best for predicting lung cancer. It achieved an accuracy of 98.38%, a precision of 98.38%, and an F1 score of 99.17%. The accuracy, precision, recall, and F1 scores for all models are presented in Table 3:



```
                           Accuracy  Precision    Recall  F1-Score
Logistic Regression        0.967742   0.983333  0.983333  0.983333
Gaussian Naive Bayes       0.951613   0.983051  0.966667  0.974790
Bernoulli Naive Bayes      0.951613   0.983051  0.966667  0.974790
Support Vector Machine     0.967742   0.983333  0.983333  0.983333
Random Forest              0.967742   0.983333  0.983333  0.983333
Decision Trees             0.967742   0.983333  0.983333  0.983333
Gradient Boosting          0.951613   0.983051  0.966667  0.974790
K-Nearest Neighbours       0.935484   0.982759  0.950000  0.966102
Extreme Gradient Boosting  0.983871   0.983607  1.000000  0.991736
Extra Tree                 0.967742   0.983333  0.983333  0.983333
ADA Boost                  0.983871   0.983607  1.000000  0.991736
Cat Boost                  0.967742   0.983333  0.983333  0.983333
Multi-layer Perceptron     0.967742   0.983333  0.983333  0.983333
```

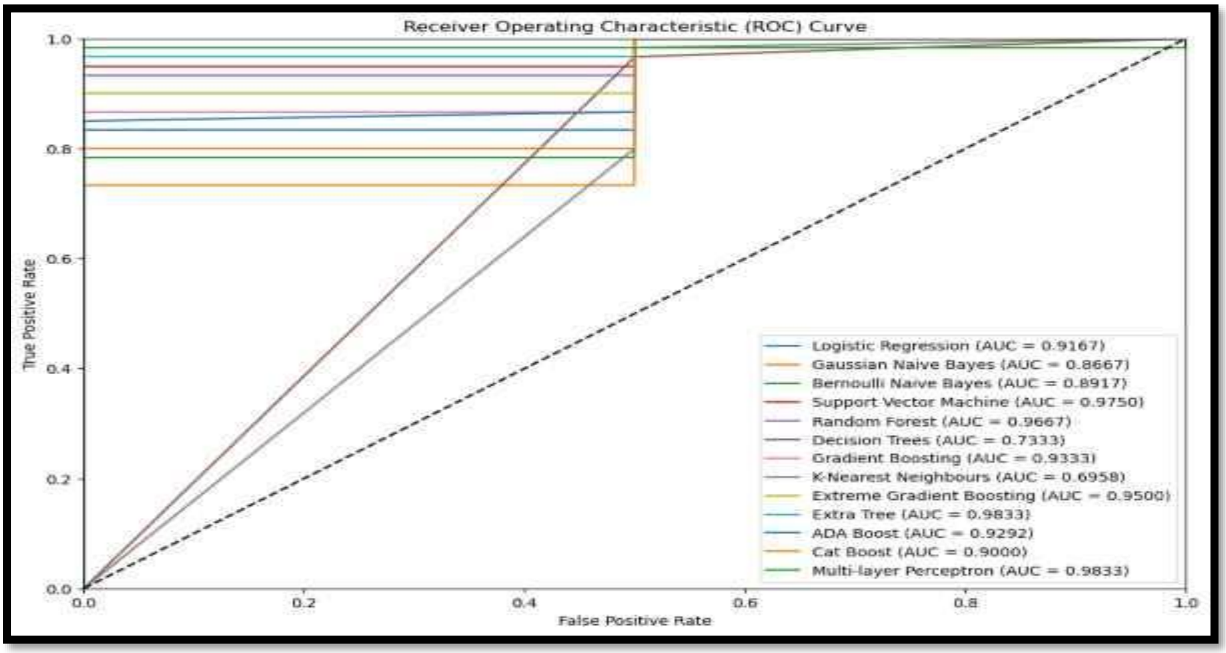
Additionally, we visualized the data using graphs to gain a clearer understanding of the models' performance:





The prediction models for lung cancer can leverage Extreme Gradient Boosting algorithms. This approach brings new insights into the design of learning algorithms. Even with the use of multiple base classifiers, the Extreme Gradient Boosting model rarely overfits and efficiently reduces the exponential loss function by building a stepwise additive model.

7. Result: In this research, we assessed 13 machine learning algorithms utilizing 15 easily accessible clinical features derived from electronic medical records. By comparing the predictive performance of individual clinical features against machine learning algorithms that employed combinations of these features, we found that the Extreme Gradient Boosting model significantly outperformed the others. This research aimed to evaluate multiple models based on different algorithms for predicting lung cancer. We found that the Extreme Gradient Boosting model achieved an impressive accuracy of 98.38%, a precision of 98.38%, and an F1 score of 99.17% in forecasting the onset of lung cancer.



The Extreme Gradient Boosting model (98.38% accuracy, ROC–AUC score of 95.0%) stand out as the leading models, outperforming the others.

I believe this study is the first to combine traditional laboratory indicators with easily accessible clinical data from electronic medical records within an Extreme Gradient Boosting model to predict lung cancer.

8. Conclusion:

This paper presents an in-depth assessment of 13 machine learning methods for forecasting lung cancer onset. The Extreme Gradient Boosting model demonstrates the highest accuracy and reliability in its predictions. Nevertheless, issues like model interpretability and data imbalance must be resolved to enable the use of these models in clinical settings. Future studies should prioritize creating more interpretable models and refining approaches to manage imbalanced datasets. In the end, leveraging machine learning for lung cancer prediction has considerable potential to boost early detection and enhance patient outcomes.

References:

[1]. Cochran AJ. 1997, Prediction of outcome for patients with cutaneous melanoma. Pigment Cell Res 1997;10:162–7.

[2]. T. Hanai, Y. Yatabe, Y. Nakayama, et al., 2003,Prognostic models in patients with nonsmall-cell lung cancer using artificial neural networks in comparison with logistic regression, Canc. Sci. 94 (5) 473–477.

[3]. P. Bach, M. Kattan, M. Thornquist, et al.,2003, Variations in lung cancer risk among smokers, J. Natl. Cancer Inst. 95 (6) 470–478.

[4]. A. Cassidy, S. Duffy, J. Myles, T. Liloglou, J. Field, 2006,Lung cancer risk prediction: a tool for early detection, Int. J. Canc. 120 (1) 1–6

[5]. Cruz JA, Wishart DS. 2007,Applications of machine learning in cancer prediction and prognosis. Cancer Inform ;2:59–77.

[6]. Kolda TG, Bader BW,2009, Tensor decompositions and applications. SIAM Rev 2009;51:455–500.

[7]. Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, et al.,2010, International network of cancer genome projects. Nature ;464:993–8.

[8]. Chi EC, Kolda TG,2012 On tensors, sparsity, and nonnegative factorizations. SIAM J Matrix Anal Appl ;33:1272–99.

[9]. C. Jeong, S. Jeong, S. Hong, et al.2012, Nomograms to predict the pathological stage of clinically localized prostate cancer in Korean men: comparison with Western predictive tools using decision curve analysis, Int. J. Urol. 19 (9) 846–852.

[10]. A.M. Deal, M.I. Milowsky, 2013,Tools to improve clinical trial design in urothelial cancer, Cancer 119 (16) 2950–2952.

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

[12]. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al.,2013, The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J Digit Imaging ;26:1045–57.

[13]. K. Ahmed, A. Al-Emran, T. Jesmin, R. Mukti, Z. Rahman, F. Ahmed, 2013, Early detection of lung cancer risk using data mining, Asian Pac. J. Cancer Prev. APJCP 14 (1) 595–598.

[14]. Luo Y, Sohani AR, Hochberg EP, Szolovits P.,2014, Automatic lymphoma classification with sentence subgraph mining from pathology reports. J Am Med Inform Assoc ;21:824–32.

[15]. P. Ramachandran, N. Girija, T. Bhuvaneswari, 2014,Early detection and prevention of cancer using data mining techniques, Int. J. Comput. Appl. 97 (13) 48–53.

[16]. P. Thangaraju, G. Barkavi, T. Karthikeyan,2014, Mining lung cancer data for smokers and NonSmokers by using data mining techniques, Int. J. Adv. Res. Comput. Commun. Eng. 3 (7) 7622–7626.

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

[18]. Pavlopoulou A, Spandidos DA, Michalopoulos I.,2015, Human cancer databases (review). Oncol Rep;33:3–18.

[19]. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J 2015;13:8–17.

[20]. Luo Y, Xin Y, Hochberg E, Joshi R, Uzuner O, Szolovits P.,2015, Subgraph augmented non-negative tensor factorization (SANTF) for modeling clinical narrative text. J Am Med Inform Assoc;22:1009–19.

[21]. S. Tiwari, N. Walia, H. Singh, A. Sharma,2015, Effective analysis of lung infection using fuzzy rules, Int. J. Bio-Sci. Bio-Technol. 7 (6) 85–96.

[22]. M. Saii, A. Mia,2015, Lung detection and segmentation using marker watershed and laplacian filtering, Int. J. Biomed. Eng. Clin. Sci. 1 (2) 29–42.

[23]. D.S. Ettinger, D.E. Wood, L.D. Aisner, W. Akerley, J. Bauman, A.L. Bazhenova, et al., 2016,Non–small cell lung cancer, version 1. J. Natl. Compr. Canc. Netw. (2016 October 14).

[24]. Luo Y, Wang F, Szolovits P.,2016, Tensor factorization toward precision medicine. Brief Bioinform;18:511–4

[25]. M. Billah, N. Islam,2016, An early diagnosis system for predicting lung cancer risk using adaptive neuro fuzzy inference system and linear discriminant analysis, J. MPE Mol. Pathol. Epidoemiol. 1 (3) 1–4.

[26]. T. Christopher, J. Jamera,2016, Study of classification algorithm for lung cancer prediction, Int. J. Innovat. Sci. Eng. Technol. 3 (2) 42–49.

[27]. A. Mishra, B. Ratha,2016, Study of random tree and random forest data mining algorithms for microarray data analysis, Int. J. Adv. Elcetric. Comput. Eng. 3 (4) 5–7.

[28]. Chiefs of Ontario,2017, Lung Cancer in First Nations People in Ontario. Ontario, Cancer Care Ontario and Institute for Clinical Evaluative Sciences.

[29].D.S. Ettinger, D.E. Wood, L.D. Aisner, W. Akerley, J. Bauman, A.L. Bazhenova, et al., 2017,Non–small cell lung cancer, version 1.

[30]. M. Kennedy, P. Beddy, J. Bruzzi, J. Bruzzi, J. Murray, K. O'Regan, et al.,2017, Diagnosis, Staging and Treatment of Lung Cancer (NCEC National Clinical Guideline, sixteenth ed., Department of Health, Dublin. Available at: http://health.gov.ie/nation al-patient-safety-office/ncec/national-clinical-guidelines.

[31].D. Shead, A. Corrigan, S. Kidney, L. Hanisch, R. Clarke, K. Williams,2017, Lung Cancer Screening, first ed., National Comprehensive Cancer Network, Washington.

[32]. Zeng Z, Li X, Espino S, Roy A, Kitsch K, Clare S, et al. 2017,Contralateral breast cancer event detection using natural language processing. AMIA Annu Symp Proc 2017:1885–92.

[33]. T. Manikandan, N. Bharathi, M. Sathish, V. Asokan,2017, Hybrid neuro-fuzzy system for prediction of lung diseases based on the observed symptom values, J. Chem. Pharmaceut. Sci. 8 69–76.

[34]. S. Durga, K. Kasturi,2017, Lung disease prediction system using data mining techniques, J. Adv. Res. Dyn. Control Sys. 9 (5) 62–66.

[35]. H. Bharathi, T.S. Arulananth,2017, A review of lung cancer prediction system using data mining techniques and self organizing map (SOM), Int. J. Appl. Eng. Res. 12 (10) 2190–2195.

[36]. D.E. Wood, E.A. Kazerooni, S.L. Baum, G.A. Eapen, D.S. Ettinger, L. Hou, et al.,2018, Lung cancer screening, version 3., NCCN clinical practice guidelines in oncology, J. Natl. Compr. Canc. Netw. 16 (4) (2018 Apr 1) 412–441. 31.

[37]. Chao G,Mao C,Wang F, Zhao Y, Luo Y.2018, Supervised nonnegative matrix factorization to predict icu mortality risk. Proceedings (IEEE Int Conf Bioinformatics Biomed) 2018:1189–94.

[38]. Zeng Z, Espino S, Roy A, Li X, Khan SA, Clare SE, et al.,2018, Using natural language processing and machine learning to identify breast cancer local recurrence. BMC Bioinformatics;19:498.

[39]. S. Senthil, B. Ayshwarya,2018, Lung cancer prediction using feed forward back propagation neural networks with optimal features, Int. J. Appl. Eng. Res. 13 (1) 318–325.

[40]. M. Markaki, I. Tsamardinos, A. Langhammer, V. Lagani, K. Hveem, O.D. Røe,2018, A validated clinical risk prediction model for lung cancer in smokers of all ages and exposure types: a hunt study, EBioMedicine 31 36–46.

[41]. Zeng Z, Yao L, Roy A, Li X, Espino S, Clare SE, et al.,2019, Identifying breast cancer distant recurrences from electronic health records using machine learning. J Healthc Inform Res;3:283–99.

[42]. Svoboda E.,2020, Artificial intelligence is improving the detection of lung cancer. Nature ;587:S20–2.

[43]. Wang H, Li Y, Khan SA, Luo Y.,2020, Prediction of breast cancer distant recurrence using natural language processing and knowledge-guided convolutional neural network. Artif Intell Med;110:101977.

[44]. Benzekry S.,2020, Artificial intelligence and mechanistic modeling for clinical decision making in oncology. Clin Pharmacol Ther;108:471–86.

[45]. Zeng Z, Vo A, Li X, Shidfar A, Saldana P, Blanco L, et al.,2020, Somatic genetic aberrations in benign breast disease and the risk of subsequent breast cancer. NPJ Breast Cancer ;6:24.

[46].Thai AA, Solomon BJ, Sequist LV, Gainor JF, Heist RS.,2021, Lung cancer. Lancet ;398:535–54.

[47]. Tran KA, Kondrashova O, Bradley A, Williams ED, Pearson JV, Waddell N.,2021, Deep learning in cancer diagnosis, prognosis and treatment selection. Genome Med;13:152.

[48]. Li Y, Luo Y.,2021, Optimizing the evaluation of gene-targeted panels for tumor mutational burden estimation. Sci Rep;11:21072.

[49]. Bhinder B, Gilvary C, Madhukar NS,2021, Elemento O. Artificial intelligence in cancer research and precision medicine. Cancer Discov;11:900–15.

[50]. Na J, Zong N, Wang C, Midthun DE, Luo Y, Yang P, et al.,2021, Characterizing phenotypic abnormalities associated with high risk individuals developing lung cancer using electronic health records from the All of Us researcher workbench. J Am Med Inform Assoc ;28:2313–24.

[51]. L.A. Vale-Silva, K. Rohr,2021, Long-term cancer survival prediction using multimodal deep learning, in English, Sci Rep-Uk 11 (1). ARTN 1350510.1038/ s41598-021-92799-4.

[52]. E.M. Nwanosike, B.R. Conway, H.A. Merchant, S.S. Hasan,2022, Potential applications and performance of machine learning techniques and algorithms in clinical practice: a systematic review, Int. J. Med. Inf. 159, 104679, https:// doi.org/10.1016/j.ijmedinf.2021.104679.

[53]. Luchini C, Pea A, Scarpa A.,2022 Artificial intelligence in oncology: current applications and future perspectives. Br J Cancer;126:4–9.

[54]. WHO, "Cancer [Online],2022, Available: https://www.who.int/news-room/fact-sheets /detail/cancer.

[55]. S. Tomassini, N. Falcionelli, P. Sernani, L. Burattini, A.F. Dragoni,2022, Lung nodule diagnosis and cancer histology classification from computed tomography data by convolutional neural networks: a survey, Comput. Biol. Med. 146, 105691, https://doi.org/10.1016/j.compbiomed.2022.105691.

[56]. Y. Yang, L. Xu, L. Sun, P. Zhang, S.S. Farid,2022, Machine learning application in personalised lung cancer recurrence and survivability prediction, Comput. Struct. Biotechnol. J. 20, 1811–1820, https://doi.org/10.1016/j.csbj.2022.03.035.

[57]. I. Kaur, M.N. Doja, T. Ahmad,2022, Data mining and machine learning in cancer survival research: an overview and future recommendations, J. Biomed. Inf. 128, 104026, https://doi.org/10.1016/j.jbi.2022.104026.

[58]. D P Singh,2024, An Extensive Examination of Machine Learning Methods for Identifying Diabetes, Tuijin Jishu/Journal of Propulsion Technology ISSN: 1001-4055, 2024; 45: 2.

[59]. D P Singh,2024, An Extensive Analysis of Machine Learning Models to Predict the Breast Cancer Recurrence, Tuijin Jishu/Journal of Propulsion Technology ISSN: 1001-4055, 2024; 45: 2.

[60]. D P Singh,2024, A Notable Utilization Of Machine Learning Techniques In The Healthcare Sector For Optimizing Resources and Enhancing Operational Efficiency, European Journal of Biomedical and Pharmaceutical sciences, ISSN:2349-8870,Volume: 11,Issue: 7,Page:212-224, 2024,http://www.ejbps.com