RESEARCH ARTICLE

Editorial Process: Submission:06/01/2025 Acceptance:01/27/2026 Published:05/18/2026

Hybrid Modelling of Pulmonary Cancer Risk Prediction Using Classical Algorithms to Modern Machine Learning Techniques

Seeta Devi[1]*, Roshan Yadav[2], Joyce Robert Mathivanan[3], Sonopant Joshi[1], Bhagyashree Jogdeo[4]

Abstract

Background: Despite significant advancements in oncology, early diagnosis of pulmonary cancer poses a clinical challenge, thus making it a leading cause of cancer-related mortality and a focal point for the development of data-driven prediction models. The objective of the study was to predict pulmonary cancer using hybrid machine learning models. **Methods:** This study presents a comprehensive review of machine learning (ML) algorithms to facilitate early prediction of pulmonary carcinoma using electronic medical records (EMRs) data. The dataset comprising 1000 patient records and 25 predictor variables, was subjected to rigorous pre-processing, including label correction, multicollinearity assessment, and dimensionality reduction. Eighteen statistically significant features, encompassing symptoms, lifestyle factors, and environmental exposures were identified through variance inflation factor (VIF) analysis and chi-square testing. Multiple ML models, including Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), and Deep Learning (DL) classifiers, were trained and evaluated using precision, recall, F1 score, specificity, and AUC metrics. **Results:** The chi-square test revealed that age (χ^2=44.187, p<0.001), passive smoking (χ^2=752.960, p<0.001), obesity (χ^2=712.088, p<0.001), smoking (χ^2=671.006, p<0.001), and symptoms like coughing blood (χ^2=818.669, p<0.001) were significantly associated with pulmonary Carcinoma. The performance metrics indicate that most basic and ensemble models, including DT, SVM, LR, KNN, AdaBoost, and RF, achieved perfect scores (accuracy, precision, recall, F1, AUC = 1.000), demonstrating optimal classification. DL and SVM Bagging showed 97% accuracy, while NN and MLP performed well with accuracy above 96%, though slightly less than the ensemble models. **Conclusion:** These findings accentuate the potential of ML, especially SVM, for early prediction of pulmonary carcinoma using structured EMR data. These findings support the integration of ML-based tools into clinical workflows, supporting data-driven, personalized cancer screening and decision-making in health care.

Keywords: Lung cancer- Pulmonary carcinoma- Machine learning- Deep learning- Ensemble- Prediction

Asian Pac J Cancer Prev, **27 (5)**, 1641-1654

Introduction

Globally, pulmonary carcinoma is the foremost widespread and fatal malignancies that claimed 1.8 million lives in 2022, 18.7% of all cancer deaths globally, emphasizing the critical importance of early, accurate detection for improving survival rates [1, 2]. Traditional cancer diagnostics including radiography, CT scans, and histopathology are though effective, face significant drawbacks including high costs, time constraints, and subjective interpretation. Over the past decade, artificial intelligence applications expanded dramatically, driven by advanced algorithms, increased computing power, and better organized datasets [3]. Over the past two decades, Artificial Intelligence (AI) and Machine Learning (ML) algorithms have emerged in enhancing human analytical capabilities with inconsistent data and reliable decisions. These technologies now permeate virtually every aspect of modern life. Various ML algorithms are specifically designed to categorize, generate predictions, or optimize raw data. Although the suggested methods are not universally applicable, several algorithms have demonstrated effectiveness in data prediction [4].

ML algorithms analyse extensive datasets including medical images, patient records, and genomic profiles to derive valuable insights through pattern recognition. In lung cancer, convolutional neural networks (CNNs) are primarily used to analyse radiological images, such as CT scans, for the early detection of suspicious lesions or nodules. When combined with image segmentation

*[1]Symbiosis College of Nursing (SCON), Symbiosis International (Deemed University), Maharashtra, India. [2]B. Tech. in Artificial Intelligence and Machine Learning; Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune, Maharashtra, India. [3]Tutor, Symbiosis College of Nursing (SCON), Symbiosis International (Deemed University), Maharashtra, India. [4]Department of Child Health Nursing, Bharati Vidyapeeth (Deemed to be University), Pune. *For Correspondence: drseetadevi1981@gmail.com*

algorithms, CNN can precisely locate and measure lung nodules [5]. These advanced models outperform traditional image processing techniques because they learn hierarchical feature representations directly from raw data, reducing dependence on manually crafted features. ML is also making significant healthcare impacts in three key areas: clinicians benefit from swift, precise image interpretation; health systems improve workflows and reduce medical errors; and patients gain the ability to manage their personal health data more effectively [6]. Google's deep learning model exemplifies this potential, achieving radiologist-level accuracy in pulmonary cancer detection from CT scans [7]. Furthermore, hybrid models that integrate image analysis with electronic health record data significantly improve prediction accuracy, particularly among high-risk populations [8].

Despite the advances, there are several challenges, especially in healthcare to know what the AI algorithm is doing; it should not be a 'black box'. Explainable AI (XAI) enables healthcare providers and patients to comprehend the reasoning behind AI-driven decisions, thereby enhancing confidence in the system's output or diagnostic prediction. It is also important that AI works according to privacy laws, is free from bias, and does not produce toxic language (in case of a medical chatbot) [9]. Several of these issues arise from the basic aim in healthcare to deliver tailored predictions using data produced and handled by medical infrastructures, where the main objective of data gathering is to aid clinical care, not future analysis. Transparency and big data use in precision medicine are essential [10]. Unsupervised ML models offer a promising approach by identifying patterns in datasets without requiring outcome information, making them valuable tools for detecting anomalies and potential fraud [11].

This study explores the effectiveness of various machine learning (ML) algorithms in predicting lung cancer, focusing on model accuracy, sensitivity, and specificity. By comparing basic ML models, deep learning techniques, and ensemble methods, the study aims to identify the optimal approaches for pulmonary cancer prediction using electronic medical record (EMR) data. The originality of the experiment is in its methodology, as it introduces inter-variable correlation analysis and multi-collinearity testing to inform the decision on what variables to include in the predictor model training. Besides, it uses ML models of different complexity to fit the data accordingly, as well as compares the training time to balance performance and latency. The dataset includes 23 features representing risk factors and symptoms associated with lung cancer. The technical words used in this paper are represented in abbreviations, as shown in Table 1.

Machine Learning Algorithms in Pulmonary Cancer Prediction

Although pulmonary carcinoma has been recognized as the deadliest cancer, with early detection being crucial for a good prognosis and efficient treatment. Recent advancements in deep learning have shown great importance in pulmonary cancer diagnosis particularly in detecting nodules, classifying lesions, and predicting outcomes [12] from chest X-rays (CXRs) and CT scans [13]. These techniques can improve diagnostic accuracy, reduce observer variability, and enable earlier intervention, leading to better patient outcomes [14]. As clinical repositories grow both image and text data play crucial roles in diagnosis. Studies on pulmonary carcinoma detection focuses on clinical symptoms, therapeutic choices, and the application of AI, image analysis, and ML methods. Researchers have applied NN, SVM, DT, RF, Gradient Boosting Decision Trees (GBDT), k-NN, ANN to enhance prediction accuracy, predictive power, and robustness in handling complex data interactions [15-17]. Comparable studies with supervised learning models, deep learning models and CNN have shown refined diagnostic accuracy, both in classification and regression tasks, especially applied to dataset such as SEER (Surveillance,

Table 1. Abbreviations Used in This Document

Abbreviation	Technical Term
CT	Computed Tomography
AI	Artificial Intelligence
ML	Machine Learning
CNN	Convolutional Neural Network
xAI	Explainable AI
EMR	Electronic Medical Record
CXR	Check X-Ray
NN	Neural Network
SVM	Support Vector Machine
DT	Decision Tree
RF	Random Forest
GBDT	Gradient Boosting Decision Trees
KNN	K-Nearest Neighbours
ANN	Artificial Neural Network
SEER	Surveillance, Epidemiology and End Results
LDA	Linear Discriminant Analysis
LR	Logistic Regression
XGBoost	Extreme Gradient Boosting
AdaBoost	Adaptive Boosting
NB	Naïve Bayes
ROC	Receiver Operating Characteristics
AUC	Area Under Curve
VIF	Variation Inflation Factor
GridSearchCV	Grid Search Cross Validation
MLP	Multi-Layered Perceptron
MNB	Multinomial Naïve Bayes
RBF	Radial Basis Function
EHR	Electronic Health Record
LBP	Local Binary Pattern
LASSO	Least Absolute Shrinkage and Selection Operator
SUHRC	Symbiosis University Hospital and Research Center

Epidemiology, and End Results) [18, 19].

Recent studies highlighted the growth of ML algorithm such as NN, SVM, Extreme Gradient Boosting, RF and LDA, for lung disease prediction [20], k-NN, DT, and NB, along with the deep learning model Inception V3 [21]. These models were assessed using evaluation measures such as Accuracy, Precision, Recall and F1 Score and Area Under the Receiver Operating Characteristic Curve (AUC) [20] and confusion metrics. Confusion metrics was used to assess predicted outcomes [22]. Similarly, ML classifiers like LR, LDA, NB, k-NN, SVM, DT, RF, XGBoost, AdaBoost and ANN was applied, and the model performance was employed to evaluate using the receiver operating characteristic curve (ROC) and accuracy, the Variance Inflation Factor (VIF) was calculated to ensure no multicollinearity among variables [23]. The AUC serves as a crucial indicator of a ML model's capacity to differentiate pulmonary cancer from non-cancer cases. [24]. These algorithms proved to be highly applicable to large, real-world temporal datasets reinforcing the potential clinical risk prediction [25].

As an example, [25] made a 3-year risk predictor with a high-dimensional LASSO regression with a large Electronic Health Record (EHR) dataset and demonstrated an AUC of 0.76 and was able to identify a stressed-population group. Other researchers have taken the advantage of publicly available symptom based datasets of Kaggle. Dritsas and Trigka [24] applied SMOTE to resolve the issue of class imbalance and realized that a Rotation Forest model was much better than other classifiers with an outstanding 99.3% AUC. In the same manner, Sinjanka et al. [26] trained a Random Forest algorithm on a Kaggle dataset to obtain a high precision of 97.9 and an F1 score of 98. The excellence of the Random Forest (RF) classifier was further demonstrated in the works of Lamba et al. [28], where using SMOTE and Bayesian optimization, the use of RF gave 96.11% accuracy and by M. Prabu [29], where RF gave 100% accuracy in identifying low, medium, and high risks of lung cancer as compared to AdaBoost, Decision Tree and SVM.

These studies had a number of limitations and gaps in research, with the main ones being connected with the quality, size, and generalizability of data. Models based on actual data, like Chandran et al., could not obtain pathologic verification of cases of cancer, had to use diagnosis codes instead, and were trained on data disproportionately represented by White, non-Hispanic patients. Research that relied on publicly available datasets had other limitations; Dritsas and Trigka [24] admitted that their model was not informed by more high-quality clinical data gathered in a hospital environment, and Sinjaka et al. [26] have stated that their research was restricted by a small and local sample size and an observational study design that does not allow making claims of causality. Likewise, the article by Lamba et al. [28] pointed out that the class balancing method based on SMOTE may introduce overfitting risk and that their small dataset prevented the generalization of their model. The research by Zhao et al. [23] also had a short sample size of only one medical facility and incomplete clinical records of all patients, and this may affect the robustness of the models.

The main advantages of our study are the anonymized and validated data gathered during a period in the hospital SUHRC, the usage of feature selection techniques and the exploration of multiple machine learning models. We used a large sample of validated data and evaluated our models on different performance metrics in order to avoid overfitting and ensure model robustness. Incorporating patient habits and symptoms is crucial to enhance the precision of pulmonary cancer prediction and diagnosis. Equally, selecting appropriate analytical approach for such multifaceted datasets is essential yet a comprehensive comparison of ML methods for pulmonary cancer prediction remain limited, indicating an opportunity for further exploration.

Materials and Methods

Data Type

The data is structured into a table with proper headers but needed data type validation and enforcement.

Ethical Approvals

The dataset comprised of twenty-five predictor variables and one outcome variable across 1000 EMR records. To pre-process the data, it was essential to statistically analyse all features, correct the labels, handle missing values, and balance the dataset. Univariate analysis enabled understand individual variable distributions, while multicollinearity was addressed using the Variance Inflation Factor (VIF), leading to the removal of highly correlated features such as "Occupational Hazards" and "Genetic Risk." Chi-square tests identified 18 statistically significant predictors ($p < 0.05$), including age, smoking, air pollution, fatigue, chest pain, and other key variables. These features were selected for model training to reduce dimensionality and avoid overfitting.

The dataset was divided into training (70%) and testing (30%) subsets, and several ML classifiers were developed with optimised hyperparameters through GridSearchCV. Model performance was evaluated using accuracy, precision, recall, specificity, F1 score, and AUC. Confusion matrices and ROC curves validated model performance across three risk levels: low, medium, and high. While most models achieved near-perfect scores, SVM emerged as optimal model due to its high accuracy, computational efficiency and prediction time, making it ideal for efficient pulmonary cancer risk prediction.

System

Since, the data is not that huge, any system with more than 4GB RAM and four CPU cores can support this training. We performed the training on a normal PC with 8GB RAM and four CPU cores. We used Python for programming and following python libraries – pandas for data preparation, numpy for data transformation, scikit-learn for machine learning, seaborn and matplotlib for data visualization and xgboost for XGBoost model.

Methodology

This study is about finding an optimized approach to dig insights from the real-world data we have, select independent predictor variables and best fit model to make predictions. For insights, we performed statistical analysis like univariate analysis, bivariate analysis (chi-square test) and multivariate analysis (VIF for multi-collinearity detection). In univariate analysis, we analysed the central measures (mean, median, mode, min, max, etc.) of each variable.

With high dimension, there is a risk of dimensionality curse of overfitting. We did to multi-collinearity to detect the redundant variables. Multi-collinearity is a statistical phenomenon in machine learning where independent variables are very much correlated with each other and thus finding the independent effect of each variable on the dependent variable becomes a difficult task. This condition causes the variance of the estimated coefficients to be inflated and causes unstable and unreliable estimation results of this model, which will affect the accuracy of predictions. VIF is a diagnostic to measure the Multi-Collinearity. It is an indication of how much the estimate of a regression coefficient will be inflated by the correlation between the independent variable and the other independent variables in the model. A VIF of one indicates that there is no multi-collinearity, and higher values indicate that there is problematic multi-collinearity that can make it difficult to determine the individual impact of each variable on the dependent variable. If VIF score was higher than 50, we eliminated the variable. We selected 18 variables out of all after VIF calculations until we had all the VIFs below 50.

Chi-square test is a non-parametric statistical test which is used to find out if there is any statistically, significant relationship any two categorical variables by comparing observed frequencies to expected frequencies. Given that we wanted to quantify the contribution of the predictor variables to determining the outcome, we performed chi-square test of all the categorical predictor variables with outcome variable. Those affecting the outcome significantly were picked and the rest were discarded based on p-value. Before performing the test, we selected the value for p (level of significance) to be 0.05. If p-value < 0.05, suggests that the variable is significantly related to the outcome variable. This suggests that the variable is a potentially useful predictor and should be chosen to be included in the model. Conversely, a high p-value is an indication of a lack of significant association, and hence the candidate for discarding the variable by considering it may not contribute meaningfully to the model.

We selected three buckets of ML models i.e. basic, ensemble and deep learning based. The fundamental machine learning algorithms are basic ML models. They are also usually not as complex and computationally intensive as ensemble or deep learning techniques. Although they work well in most cases, they may be restricted by very large variance (overfitting) or very large bias (underfitting) when applied to complex data sets. We used decision tree (DT), support vector machine (SVM), logistic regression (LR), k-nearest neighbours (k-NN), linear discriminant analysis (LDA) and Multinomial Naïve Bayes (MNB).

Ensemble techniques combine predictions of multiple individual, so-called base or weak models to produce one that is stronger and more accurate. The principle behind it is that a heterogeneous group of learners makes a decision that is better than that of an individual learner. Ensemble models can be categorized into the following: Bagging, Voting, Boosting and Stacking. We used AdaBoost and XGBoost as boosting models, a Soft Voting Classifier and a Hard Voting Classifier as voting models, Random Forest (RF), k -NN Bagging, SVM Bagging as bagging models, and a Stacked Classifier as a stacking model in our study. The model of bagging works by creating various subsamples by bootstrapping, training a different model on each subset and later combining the predictions using techniques such as voting or averaging. Voting models are used to train a group of learners with the same data and the synthesized output is calculated with the help of voting. In Hard Voting the majority decision determines the ultimate prediction in contrast to Soft Voting, whereby the predicted probabilities of all the constituent models are combined by some method, usually by taking a weighted average and the result is based on these weighted probabilities. Stacking is the act of training multiple base models and passing their predictions through a meta-model which generates the final predictions. In our study, we combined LR, SVM, DT and AdaBoost (our best performing models) in Soft Voting and Hard Voting Classifiers. In stacking, we used these same algorithms as candidate models and LR as meta-model.

We have performed a Grid-Search Cross-Validation process to find the best hyperparameters of the different models before undertaking the training of the models. Hyperparameters are those parameters that are predefined before the training process begins and the parameters are not learned with the empirical data. These parameters control the behavioural dynamics of a model and the general behaviour of a model. GridSearchCV carries out an exhaustive search over a given "grid" of hyperparameter values and tunes a machine-learning model according to this grid. Taking a model and a dictionary of hyperparameters with lists of possible values, GridSearchCV will produce all combinations of possible combinations of values. It also evaluates the model performance of each combination using a cross-validation scheme, or training and validating on different subsets of the dataset. When all the evaluations have been done, the GridSearchCV will choose the hyperparameter combination that yielded the best average performance on all the cross-validation folds and will then retrain the model with these best parameters on the full training set.

In medical research, the analysis of machine-learning models typically uses a set of performance metrics, typically a set of comprehensive performance measures, as opposed to just accuracy, since the consequences of misclassification may be disastrous. Accuracy measures the consistency of positive diagnoses and thus helps in curbing the negative consequences of false positives, including false anxiety of patients and unnecessary medical treatment. Recall or sensitivity is essential in

preventing false negatives; it guarantees that true cases of disease are not mistaken and it is particularly necessary in severe disease cases where false negativity could be fatal. Specificity, in its turn, aims at the proper detection of negative cases, which will make it easier to rule out diseases with high confidence and minimize the number of false alarms in healthy people. The F1 score combines both accuracy and recall into a single balanced score, thus providing a more reliable measure of performance than accuracy, which can be deceptive in the presence of class imbalance that is often the case with medical data, including with rare diseases. Lastly, models training time was also considered.

Results

Univariate Analysis

In this retrospective study, 59.8% and 40.2% of the records were from male and female adults respectively, with most individuals (57.4%) aged between 24 and 42. Air pollutions affected 37.5% of subjects, with 78.67% at high risk of developing cancer. Major causes identified were dust and alcohol including other factors such as workplace exposures, hereditary predisposition, underlying pulmonary conditions, inadequate nutrition, overweight, active and passive smoking (nearly 100% of heavy passive smokers were high risk). Key symptoms were chest pain, haemoptysis, tiredness, unintentional weight reduction, dyspnoea bronchial wheeze, dysphagia, digital clubbing, and recurrent upper respiratory infections, which collectively served as predictors for assessing an individual's risk of lung cancer.

Selected Features (Multivariate and Bivariate Analysis)

An inter-variable correlation analysis found many highly correlated features with a threshold of 0.8, revealing significant multicollinearity using Variance Inflation Factor (VIF). Table 2 provides a comprehensive overview of the 23 input feature variables, features like Occupational Hazards and Genetic Risk had VIF scores greater than 100 while Chronic Lung Disease, Alcohol Use, and Dust Allergy had VIF scores around 50. After removing these features, the VIF of each remaining feature was around 30.

Primarily data was analysed based on the VIF scores based on multicollinear features. Subsequently, a chi-square test determined predictors that were substantially associated with outcome variable (p-value < 0.05). Table 3 reveals 18 selected features such as coughing blood, passive smoker, obesity, smoking, imbalanced diet, chest pain, fatigue, air pollution, and age which were critical features for predicting lung cancer. Non-significant correlation was excluded. This feature selection reduced the dataset's dimensionality, enhanced model performance thus alleviating the risk of overfitting.

Performance of the ML Models

Table 4 shows the test set performance of all the ML models, (30% of processed data), assessing the generalizability. Additionally, accuracy, metrics for instance precision, recall, F1 score, specificity, and AUC score was measured. Training scores were also measured to detect overfitting, as disparities in training score and testing score and the accuracy on the test dataset indicates overfitting.

To evaluate accuracy, the proportion of correctly predicted instances out of the total number of predictions was calculated. However, accuracy does not guarantee the overall performance or generalizability of a classifier. Most of the classifiers tested achieved 100% accuracy, except for LDA (96.9%), MNB (73.4%), SVM Bagging (98.1%), NN (97.3%), and MLP (98.3%).

The precision metric quantifies the ratio of true positive predictions to the total number of instances classified as positive. In multi-class classification, it is calculated individually to each class. The value presented is the average of the micro and macro averages of the precision, recall, and F1 score metrics. Classifiers with 100% accuracy also achieved 100% precision. Other models, such as LDA (97.4%), MNB (72.3%), SVM Bagging (97.7%), NN (95.9%), and MLP (96.7%), did not achieve 100% precision.

Recall, or sensitivity, measures the proportion of true positive predictions among all actual positive instances. It is crucial in scenarios where overlooking positive cases carries serious implications. The classifiers that achieved 100% accuracy and 100% precision also attained 100% recall.

Table 2. VIF Scores before and after Removal of Multicollinear Features

S.No	Feature	Before Removing VIF	After Removing VIF
1.	Age	8.879	8.392
2.	Gender	7.229	6.124
3.	Air Pollution	18.73	16.314
4.	Alcohol use	53.488	Removed
5.	Dust Allergy	50.521	Removed
6.	Occupational Hazards	124.714	Removed
7.	Genetic Risk	112.607	Removed
8.	Chronic Lung Disease	58.05	Removed
9.	Balanced Diet	38.343	32.369
10.	Obesity	42.432	27.823
11.	Smoking	16.088	13.54
12.	Passive Smoker	33.806	27.372
13.	Chest Pain	39.991	20.861
14.	Coughing of Blood	38.681	29.377
15.	Fatigue	14.379	10.563
16.	Weight Loss	15.035	13.605
17.	Shortness of Breath	20.945	15.575
18.	Wheezing	10.631	8.871
19.	Swallowing Difficulty	11.95	8.144
20.	Clubbing of Fingernails	10.794	7.528
21.	Frequent Cold	11.774	10.209
22.	Dry Cough	10.309	9.784
23.	Snoring	8.787	7.647

Table 3. Chi-Square Test Results of the Factors Affecting the Level of the Lungs Cancer

No.	Variable Name	Variable Type	Concept Codes	Frequency	chi-square	p-value
			Socio-demographic			
1	Age (years)	Polynomial	14-73 (Range)	-	44.187	<0.001
2	Gender	Binomial	1 (Female)	598	4.668	0.097
			2 (Male)	402		
			Causes			
3	Passive Smoker	Polynomial	1	60	752.96	<0.001
			2	284		
			3	140		
			4	161		
			5	30		
			6	30		
			7	187		
			8	108		
4	Obesity	Polynomial	1	70	712.088	<0.001
			2	140		
			3	193		
			4	191		
			5	20		
			6	30		
			7	356		
5	Smoking	Polynomial	1	181	671.006	<0.001
			2	222		
			3	172		
			4	59		
			5	10		
			6	60		
			7	207		
			8	89		
6	Balanced Diet	Polynomial	1	40	588.394	<0.001
			2	231		
			3	173		
			4	61		
			5	40		
			6	159		
			7	296		
7	Fatigue	Polynomial	1	110	518.9	<0.001
			2	211		
			3	212		
			4	180		
			5	89		
			6	50		
			7	0		
			8	109		
			9	39		
8	Air Pollution	Polynomial	1	141	518.632	<0.001
			2	201		
			3	173		
			4	90		
			5	20		
			6	326		
			7	30		
			8	19		
		Symptoms				
9	Coughing of Blood	Polynomial	1	71	818.669	<0.001
			2	121		
			3	171		
			4	172		
			5	50		
			6	49		
			7	187		
			8	119		
			9	60		

Table 3. Continued

No.	Variable Name	Variable Type	Concept Codes	Frequency	chi-square	p-value
			Symptoms			
10	Chest Pain	Polynomial	1	80	524.49	<0.001
			2	181		
			3	153		
			4	191		
			5	10		
			6	40		
			7	296		
			8	30		
			9	19		
11	Shortness of Breath	Polynomial	1	80	330.81	<0.001
			2	243		
			3	140		
			4	90		
			5	87		
			6	201		
			7	89		
			8	0		
			9	70		
12	Clubbing of Fingernails	Polynomial	1	131	257.908	<0.001
			2	240		
			3	100		
			4	220		
			5	120		
			6	10		
			7	40		
			8	59		
			9	80		
13	Weight Loss	Polynomial	1	121	206.667	<0.001
			2	280		
			3	150		
			4	60		
			5	100		
			6	49		
			7	230		
			8	10		
14	Wheezing	Polynomial	1	149	201.426	<0.001
			2	240		
			3	60		
			4	163		
			5	171		
			6	68		
			7	139		
			8	10		
15	Frequent Cold	Polynomial	1	139	192.713	<0.001
			2	192		
			3	230		
			4	180		
			5	20		
			6	170		
			7	69		
16	Dry Cough	Polynomial	1	119	152.03	<0.001
			2	251		
			3	101		
			4	141		
			5	131		
			6	89		
			7	168		

Table 3. Continued

No.	Variable Name	Variable Type	Concept Codes	Frequency	chi-square	p-value
			Symptoms			
17	Swallowing Difficulty	Polynomial	1	221	113.074	<0.001
			2	160		
			3	90		
			4	189		
			5	110		
			6	91		
			7	29		
			8	110		
18	Snoring	Polynomial	1	170	91.748	<0.001
			2	300		
			3	211		
			4	131		
			5	139		
			6	39		
			7	10		

Table 4. Performance Metrics of Machine Learning Models Evaluated on the Test Dataset

No.	Model	Training Score	Accuracy	Precision	Specificity	Recall	F1 Score	AUC	Time(s)
				Basic ML Algorithms					
1	DT	1	1	1	1	1	1	1	0.856
2	SVM	1	1	1	1	1	1	1	0.169
3	LR	1	1	1	1	1	1	1	4.726
4	KNN	1	1	1	1	1	1	1	2.557
5	LDA	0.969	0.973	0.974	0.986	0.973	0.973	0.98	0.291
6	MNB	0.734	0.727	0.723	0.865	0.724	0.723	0.8	0.15
				Ensemble ML Algorithms					
7	AdaBoost	1	1	1	1	1	1	1	0.287
8	Soft Voting	1	1	1	1	1	1	1	0.169
9	XGBoost	1	1	1	1	1	1	1	8.875
10	RF	1	1	1	1	1	1	1	37.64
11	KNN Bagging	1	1	1	1	1	1	1	16.74
12	Hard Voting	1	1	1	1	1	1	1	0.135
13	Stacking	1	1	1	1	1	1	1	0.425
14	SVM Bagging	0.981	0.977	0.977	0.988	0.976	0.977	0.98	0.303
				Deep Learning Algorithms					
15	NN	0.973	0.96	0.959	0.958	0.958	0.959	1	10
16	MLP	0.983	0.967	0.967	0.982	0.967	0.967	0.98	12.42

Specificity refers to the ratio of true negative predictions to the total number of actual negative cases. Although not commonly discussed as precision and recall, it's vital when accurately identifying negative instances is critical. For the classifiers with 100% accuracy, precision, and recall, specificity is also 100%. The F1 Score is the harmonic mean of precision and recall, providing a balanced evaluation of a classifier's performance, especially in cases of uneven class distribution. Classifiers with 100% precision and recall also achieved an F1 Score of 100%.

The Area Under the ROC Curve (AUC) measures a classifier's ability to distinguish between positive and negative instances across all classification thresholds, summarizing this performance into a single metric. Most classifiers achieved the maximum AUC score of 100%, indicating strong discriminatory power. In contrast, the Multinomial Naive Bayes (MNB) classifier recorded a lower AUC score of 80%, reflecting reduced discriminatory capability compared to other models. The remaining classifiers consistently achieved AUC scores of 98%, demonstrating excellent discriminatory performance. All algorithms were tuned to optimal hyperparameters using Grid Search CV algorithm. Table 5 displays the optimal hyperparameter settings for each model.

Figure 1 displays 3×3 confusion matrices for all classifiers, summarizing their performance across three

classes: 0 (low), 1 (medium), and 2 (high). Based on these confusion matrices, evaluation metrics such as accuracy, precision, recall, specificity, and F1 score were calculated. Most classifiers achieved 100% accuracy across all classes, while the Multinomial Naive Bayes (MNB) model showed weaker performance correctly classifying approximately 65% of the "low" and "medium" classes and 85% of the "high" class.

Figure 2 presents ROC curves for all classifiers across the three classes, using blue for class 0 (low), orange for class 1 (medium), and green for class 2 (high). In this multi-class ROC analysis, each class is treated as the positive class, while the other two are treated as negative. The ROC curves plot sensitivity (true positive rate) against the false positive rate (1 – specificity) across different classification thresholds.

Models such as LR, DT, SVC, RF, XGBoost, Bagging k-NN, Soft Voting, Hard Voting, AdaBoost, k-NN, NN, and stacking demonstrated optimal performance, achieving AUC scores of 1.00 for all classes. In contrast, MNB had the lowest performance, with AUC values of 0.90, 0.86, and 0.95 for the low, medium, and high classes, respectively.

Since most classifiers achieved perfect predictive performance, the choice of the best model was primarily based on training and prediction speed. Among the basic classifiers, SVM stood out for its efficiency and simplicity, making it the selected optimal algorithm.

Discussion

Despite substantial advancement in the field of oncology, early detection of pulmonary cancer poses significant challenges. In this study we evaluated multiple ML algorithm on EHR's, suggesting potential data driven methods for early detection of lung cancer; our findings

Figure 1. Confusion Matrices of All the Classifiers for Performance Comparison

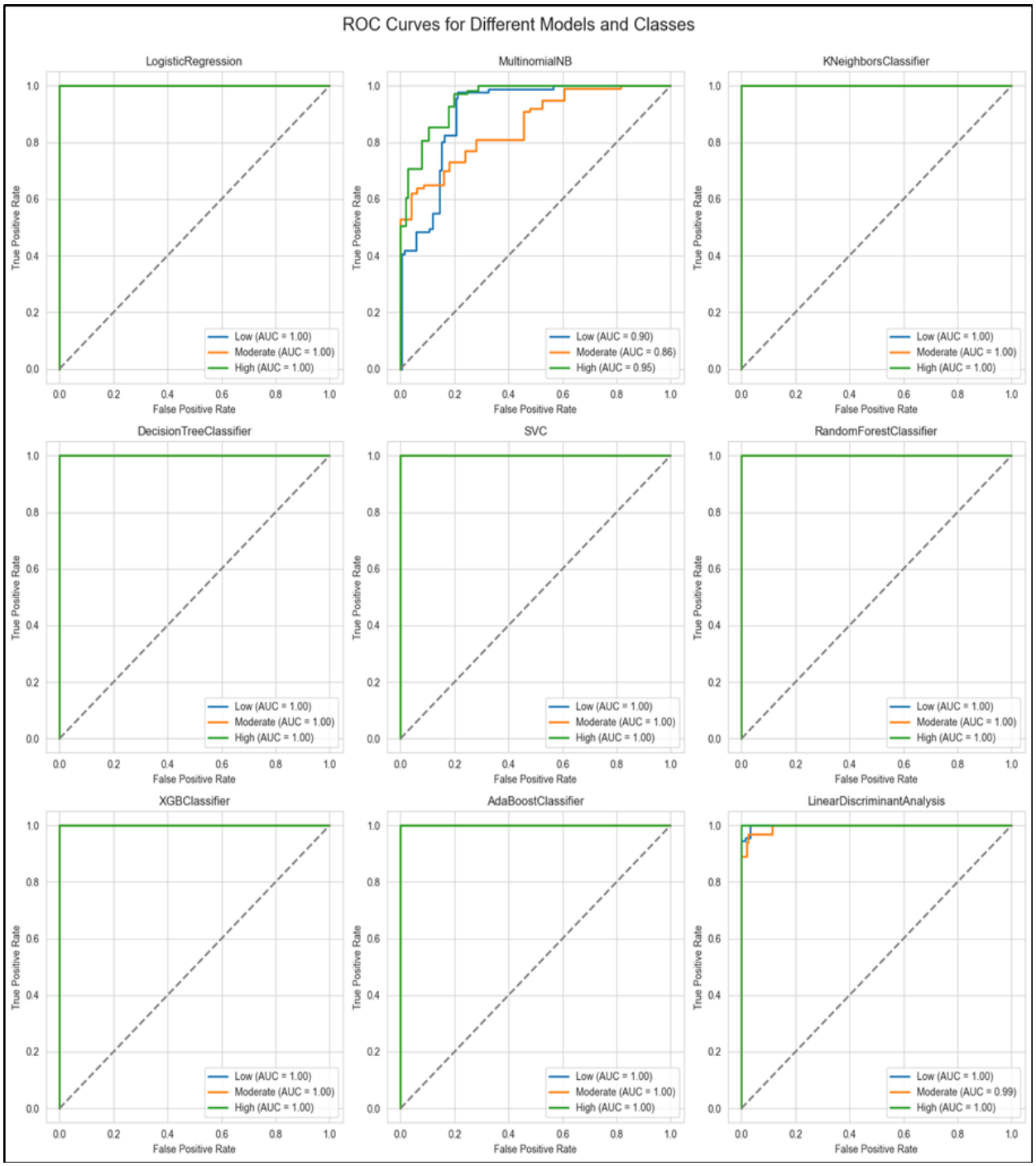

Figure 2. ROC Curves for Each Class of All the Classifiers for Performance Comparison

indicated that SVM surpassed other models. Intriguingly, similar research by Meera devi., et al, (2025), noted higher accuracy of SVM with radial basis function (RBF) kernel for classifying lung conditions using X-ray imaging data [21].

Most prevalent pulmonary cancer risk models often rely on data from clinical trials due to the absence of key predictors in clinical practice. In contrast, real-world data from EHR render practical, scalable, efficient and supplementary approach to prediction research [25]. This study utilized a dataset of 1000 EHR containing 25 predictors and one outcome variable. Similarly, Sinjanka, Y., et al, initially performed a univariate analysis, assessing essential patient information and key health indicators individually for its potential impact on pulmonary cancer prediction through key statistics and to particularly prove RF in handling complex data [26].

ML identifies patterns through algorithms to make predictions as an essential tool and advanced application for early diagnosis of lung cancer, its classification etc [27]. The ML models were evaluated for generalizability on the test dataset, along with indicators, as well as precision, recall, accuracy, and F1-score, specificity and AUC value. Lamba, R., et al. integrated multiple ML classifiers (AdaBoost, RF, and XGBoost) achieving over 95% accuracy, with RF performing best, suggesting

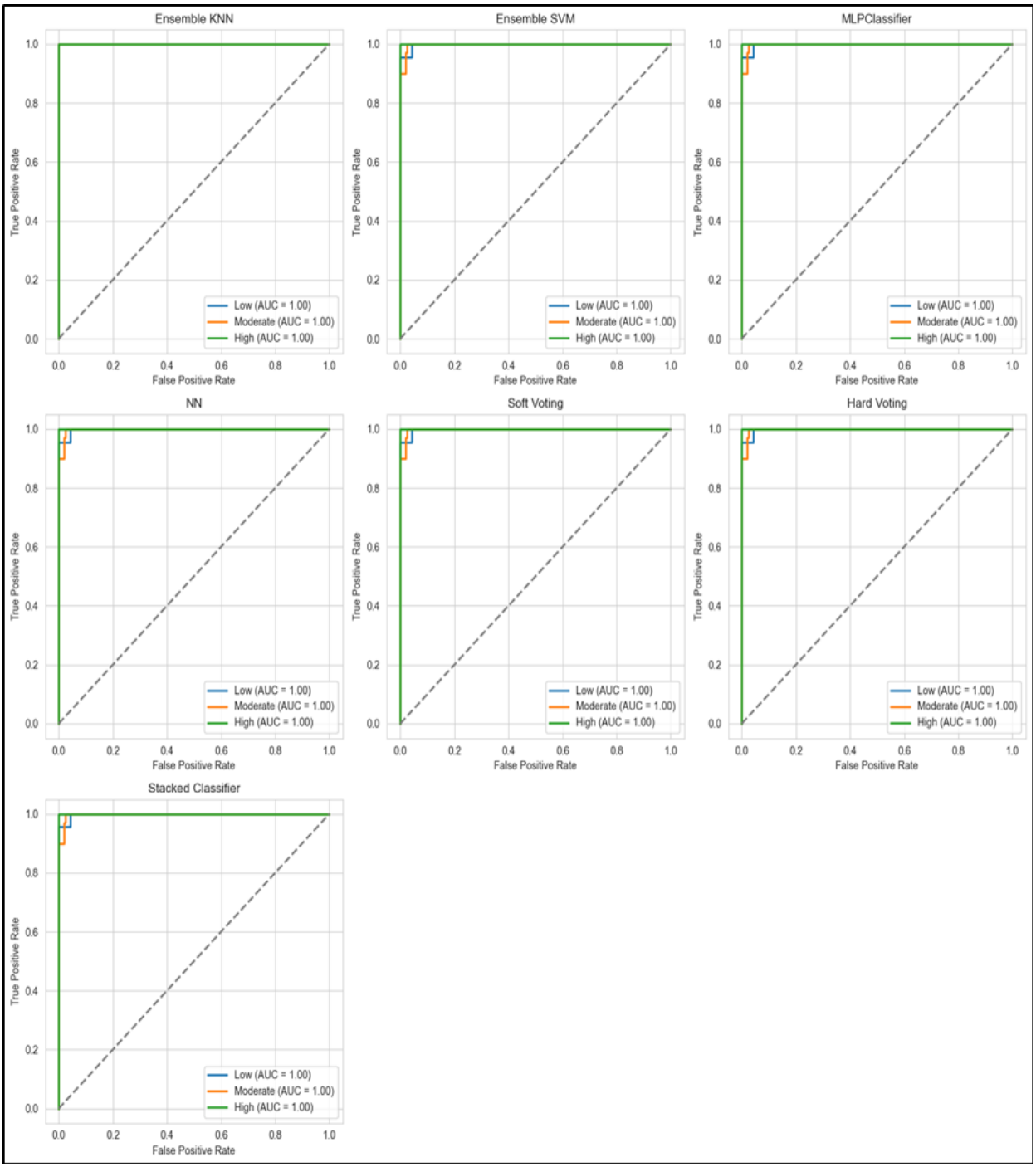

Figure 2. Continued

that testing on diverse populations can enhance model generalizability [28]. Similar results were reported that employed AdaBoost, DT, SVM, and RF [29]. Random Forest consistently outperformed due to its ability to generate multiple trees, with splits determined by feature importance. SVM, k-NN, and DT are widely used on numeric datasets, and several Kaggle datasets have been successfully trained using these models.

A recent study reported accuracies of 95.4% for SVM, 93.7% for DT, and 95.2% for k-NN [24], while another reported an accuracy of 99.2% for SVM and 90% for DT on the same dataset [30]. Similarly, while employing Local Binary Patterns (LBP) for feature extraction, substantial improvements were observed across all algorithms. Specifically, NB achieved an accuracy of 0.851, DT reached 0.912, and SVM attained an accuracy of 0.961 [31]. However, neither study performed hyperparameter tuning, which could potentially improve accuracy, and neither provided model interpretability or explanations. In contrast, our study includes systematic hyperparametric tuning, such as various k values in k-NN, fine-tuning regularization parameters (C and gamma) in SVM, and selecting optimal tree depths in decision tree models. The detailed comparison is in the Table 6.

The highlight of our study is not only the high performing model, which take minimal time for training,

Table 5. Optimal Hyperparameters of All the Classifiers Tuned Using Grid Search CV Technique

Classifier	Optimal Hyperparameters
	Basic ML Algorithms
LR	C = 1, solver = newton-cholesky
DT	criterion = gini, max_depth = 5, min_samples_leaf = 5
MBN	alpha = 0
LDA	n_components = 1, solver = svd, store_covariance = True, tol = 0.0001
KNN	leaf_size = 20, metric = minkowski, n_neighbors = 3, p = 1, weights = uniform
	Ensemble ML Algorithms
RF	max_depth = 5, min_samples_leaf = 5, n_estimators = 10
AdaBoost	base_estimator__criterion = gini, base_estimator__splitter = best, n_estimators = 10
XGBoost	colsample_bytree = 0.6, gamma = 0.5, max_depth = 4, min_child_weight = 1, subsample = 0.6
KNN Bagging	base_estimator__n_neighbors = 3, max_features = 0.7, max_samples = 0.7, n_estimators = 10
SVM Bagging	base_estimator__C = 1, base_estimator__kernel = linear, max_features = 0.9, max_samples = 0.9, n_estimators = 10
Voting Classifier	voting = soft/hard
Stacking Classifier	stack_method = auto, passthrough = True
	Deep Learning Algorithms
MLP	activation = tanh, alpha = 0.05, hidden_layer_sizes = (10, 30, 10), learning_rate = constant, solver = adam
NN	activation = relu, alpha = 0.001, hidden_layer_sizes = (64, 64), learning_rate = adaptive, solver = adam

Table 6. Comparison of Recent Studies

Study	Dataset	Feature Selection Algorithm	Best Model	Performance Metrics	Performance Measure
[24]	Lung Cancer Prediction Dataset from Kaggle	Gain Ratio and Random Forest	SVM	Accuracy, Recall and Precision	95.4% each
[25]	Licensed Optum HER	LASSO Regression	LASSO Regression	AUC	81%
[26]	Lung Cancer Prediction Dataset from Kaggle	None	Random Forest	Accuracy, Precision and Recall	97.9%, 98% and 98%
[28]	Lung Cancer Dataset from Kaggle	Correlation Analysis	Random Forest	Accuracy, Precision, Recall and Specificity	96.11%, 95.94%, 96.97% and 95.92%
[29]	Lung Cancer dataset from the UCI Machine Learning Repository.	None	Random Forest	Precision, Recall and F1-Score	100% each
Ours	Sampled Lung Cancer Data from SUHRC, Pune	Correlation Analysis, Multicollinearity (VIF)	SVM	Accuracy, Precision, Recall, Specificity, F1-Score and AUC	100% each

and inference, our feature selection approaches reduce the redundant dimensionality and ensure near-perfect model fitting. This not only improved the performance of the model but also reduced the overfitting. Not only that, the model has 100% accuracy, but there is also 100% precision, recall, specificity, and F1-score, which helps in ruling out the fears of overfitting.

In conclusion, the predictive efficiency of various ML algorithms for pulmonary cancer detection was examined systematically using EMRs. Through rigorous pre-processing and strategic feature selection, we identified eighteen critical predictors encompassing lifestyle factors, environmental variables, and clinical symptoms that demonstrate a strong association with pulmonary cancer risk. Following model training and evaluation, several classifiers particularly SVM, RF, and ensemble algorithm like XGBoost and Voting Classifiers, displayed optimal classification across critical metrics that include accuracy, precision, recall, specificity, F1 score, and AUC. Amongst them SVM emerged as the most efficient model due to its combination of computational efficiency and predictive robustness. These findings accentuate ML's transformative potentiality in improving clinical decision-making, particularly in timely identification and classification of pulmonary cancer risk. By efficiently using structured EMR data and new prediction algorithms, this study sets the groundwork for a future of data-driven, individualized cancer treatment.

Limitations and Future Scope

The current research demonstrates the robust performance of several ML models but also acknowledges the key limitations. The reliance of retrospective dataset with 1000 EMR records, though sufficient may not capture the entire variability and complexity of real-world clinical populations. Additionally, limiting the model's exposure to richer, multi-dimensional clinical data such as radiographic imaging, genomic profiles, and physician notes each of which could enhance model interpretability and prognostic capability. Future research should prioritize prospective, multi-institutional validation and explore the integration of multimodal inputs to

build more generalizable and context-aware predictive systems. Furthermore, incorporating comprehensible artificial intelligence (XAI) frameworks is pivotal for bridging the gap between algorithmic output and clinical interpretability, enabling greater trust, accountability, and ethical ML in oncology decision making.

Author Contribution Statement

Both the authors participated in discussing the findings and contributed to the final version of the manuscript.

Acknowledgements

Data availability

The datasets used and analyzed during the current study are available from, Dr. Seeta Devi., upon reasonable request.

Disclaimer

The opinions presented in this article are solely those of the authors and do not reflect the official stance of the affiliated institution.

Ethical Clearance

This study was approved by the Independent Ethics Committee of Symbiosis International (Deemed University) (SIU)

Competing Interest

The authors declare no financial or personal relationships that could have influenced the writing of this article.

References

1. World Health Organization. Global cancer burden growing amidst mounting need for services [Internet]. 2024 Feb 1 [cited 2025 May 1]. Available from: https://www.who.int/news/item/01-02-2024-global-cancer-burden-growing--amidst-mounting-need-for-services
2. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018;68(6):394–424. https://doi.org/10.3322/caac.21492
3. Ladbury C, Amini A, Govindarajan A, Mambetsariev I, Raz DJ, Massarelli E, et al. Integration of artificial intelligence in lung cancer: Rise of the machine. Cell Rep Med. 2023;4(2):100933. https://doi.org/10.1016/j.xcrm.2023.100933
4. Tuncal K, Sekeroglu B, Ozkan C. Pulmonary cancer incidence prediction using ML algorithms. J Adv Inf Technol. 2020;11(2). https://doi.org/10.12720/jait.11.2.91-96
5. Bhuiyan MS, Chowdhury IK, Haider M, Jisan AH, Jewel RM, Shahid R, et al. Advancements in early detection of pulmonary cancer in public health: a comprehensive study utilizing machine learning algorithms and predictive models. J Comput Sci Technol Stud. 2024;6(1):113-121. https://doi.org/10.32996/jcsts.2024.6.1.12
6. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44–56. https://doi.org/10.1038/s41591-018-0300-7
7. Ardila D, Kiraly AP, Bharadwaj S, Choi B, Reicher JJ, Peng L, et al. End-to-end pulmonary cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nat Med. 2019;25(6):954–961. https://doi.org/10.1038/s41591-019-0447-x
8. Maurya SP, Sisodia PS, Mishra R, Singh DP. Performance of machine learning algorithms for pulmonary cancer prediction: a comparative approach. Sci Rep. 2024;14(1):18562. https://doi.org/10.1038/s41598-024-58345-8
9. Hulsen T, Manni F. Guest Editorial: Big data and artificial intelligence in healthcare. Health Technol Lett. 2024;11(4):207–209. https://doi.org/10.1049/htl2.12086
10. Ghassemi M, Naumann T, Schulam P, Beam AL, Chen IY, Ranganath R. A review of challenges and opportunities in machine learning for health. AMIA Joint Summits Transl Sci Proc. 2020;2020:191–200.
11. Alanazi A. Using machine learning for healthcare challenges and opportunities. Informat Med Unlocked. 2022;30:100924. https://doi.org/10.1016/j.imu.2022.100924
12. Patel AN, Srinivasan K. Deep learning paradigms in pulmonary cancer diagnosis: A methodological review, open challenges, and future directions. Phys Med. 2025;131:104914. https://doi.org/10.1016/j.ejmp.2025.104914
13. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts H. Artificial intelligence in radiology. Nat Rev Cancer. 2018;18(8):500–510. https://doi.org/10.1038/s41568-018-0016-5
14. Nguyen MN. A scoping review of deep learning approaches for pulmonary cancer detection using chest radiographs and computed tomography scans. Biomed Eng Adv. 2025;9:100138. https://doi.org/10.1016/j.bea.2024.100138
15. Yang Y, Xu L, Sun L, Zhang P, Farid SS. Machine learning application in personalised pulmonary cancer recurrence and survivability prediction. Comput Struct Biotechnol J. 2022;20:1811–1820. https://doi.org/10.1016/j.csbj.2022.03.035
16. Zou Y. Pulmonary cancer prediction based on machine learning. Highlights Sci Eng Technol. 2025;124:188–191. https://doi.org/10.54097/qa56rs79
17. Patra R. Prediction of Pulmonary cancer Using Machine Learning Classifier. In: Chaubey N, Parikh S, Amin K, editors. Computing Science, Communication and Security. COMS2 2020. Communications in Computer and Information Science, vol 1235. Singapore: Springer; 2020. p. 103–110. https://doi.org/10.1007/978-981-15-6648-6_11
18. Taye MM. Understanding of machine learning with deep learning: Architectures, workflow, applications, and future directions. Computers. 2023;12(5):91. https://doi.org/10.3390/computers12050091
19. Doppalapudi S, Qiu RG, Badr Y. Pulmonary cancer survival period prediction and understanding: deep learning approaches. Int J Med Inform. 2021;148:104371. https://doi.org/10.1016/j.ijmedinf.2020.104371
20. Rajasekhar N. Performance analysis of machine learning algorithms on pulmonary cancer disease. Int J Res Stud Sci Eng Technol. 2017;4(1):41–49. ISSN 2349-4751 (Print), ISSN 2349-476X (Online).
21. Meeradevi T, Sasikala S, Murali L, Manikandan N, Ramaswamy K. Pulmonary cancer detection with machine learning classifiers with multi-attribute decision-making system and deep learning model. Sci Rep. 2025;15:8565. https://doi.org/10.1038/s41598-025-88188-w
22. Pathoee K, Rawat D, Saini DKJB. Predictive model using machine learning approach for the detection of breast cancer. Int J Comput Eng. 2024;1(2):33–37. https://doi.org/10.62527/comien.1.2.9
23. Zhao H, Su Y, Wang M, et al. The machine learning model

for distinguishing pathological subtypes of non-small cell lung cancer. Front Oncol. 2022;12:875761. https://doi.org/10.3389/fonc.2022.875761
24. Dritsas E, Trigka M. Lung cancer risk prediction with machine learning models. Big Data Cogn Comput. 2022;6(4):139. https://doi.org/10.3390/bdcc6040139
25. Chandran U, Reps J, Yang R, et al. Machine learning and real-world data to predict pulmonary cancerrisk in routine care. Cancer Epidemiol Biomarkers Prev. 2023;32(3):337–343. https://doi.org/10.1158/1055-9965.EPI-22-0873
26. Sinjanka Y, Kaur V, Musa UI, et al. ML-based early detection of lung cancer: an integrated and in-depth analytical framework. Discov Artif Intell. 2024;4:92. https://doi.org/10.1007/s44163-024-00204-6
27. Li Y, Wu X, Yang P, Jiang G, Luo Y. Machine learning for pulmonary cancer diagnosis, treatment, and prognosis. Genomics Proteomics Bioinformatics. 2022;20(5):850–866. https://doi.org/10.1016/j.gpb.2022.11.003
28. Lamba R, Rani P, Sachdeva RK, Bathla P, Kumar K, Mittal V, Joshi K. An optimized predictive machine learning model for Lung cancer diagnosis. Biomed Pharmacol J. 2025;18(March Special Edition). https://dx.doi.org/10.13005/bpj/3075
29. Prabu M. Lung cancer prediction system based on machine learning algorithms. The Bioscan. 2024;19(3):95–102. https://doi.org/10.63001/tbs.2024.v19.i03.pp95-102
30. Nair R, G. A comparative study of pulmonary cancer detection using machine learning algorithms. 2019 IEEE Int Conf Electr Comput Commun Technol (ICECCT); 2019 Mar 1-4; Coimbatore, India. p. 1-4. https://doi.org/10.1109/ICECCT.2019.8869001
31. Prakasha Raje Urs M. Machine learning approach for pulmonary cancer detection and classification–a comparative analysis. Int J Intell Syst Appl Eng. 2024;12(3):3819–3826. Available from: https://www.ijisae.org/index.php/IJISAE/article/view/6065