Lung cancer Prediction and Classification Using Machine Learning Techniques

Abstract

Lung cancer is one of the leading causes of death for both men and women in every country. Lung cancer has a significant fatality rate owing to its poor prognosis. Using image recognition and data analytics, the computing industry is becoming completely automated, as is the medical industry. This research intends to analyze the accuracy ratio of five classifiers, including Random Forest (RF), Naive Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT), and Logistic Regression, in order to detect lung cancer early and maybe save many lives. The majority of the informational indices included in this study were obtained from UCI datasets for lung cancer patients. The primary topic of this research is the execution analysis of the classification algorithm's precision. The testing results indicate that Support Vector Machine (SVM) achieves the highest success rate of 93%.

Introduction

Lung cancer is the second most common type of cancer. It affects one out of every five men and one out of every nine women. In the past few years, the number of men with lung cancer has gone down slightly, but the number of women with lung cancer has gone up a lot. In 1940, the disease was found in only seven out of every 100,000 women. Today, that number is 42. All of the statistics that are available say that smoking is what causes smoking. One person who knows a lot about the subject says that the number of cigarettes smoked every day affects how long it takes to get cancer. But data show that the risk goes down if you stop smoking. The two most common types of lung cancer are small cell lung cancer (SCLC), also called "oat cell cancer" because the cancerous cells look like oat grains, and non-small cell lung cancer (NSCLC) (NSCLC). The severity of the disease and the treatment options depend on the type of tumor found. Many types of lung cancer start and spread quickly, and the lungs are very important organs. This shows how important it is to find and treat lung cancer quickly, often by having the tumor surgically removed. Diseases like cholera, chikungunya, and cancer are becoming more common along with the rapid growth of the world's population. Cancer is becoming more and more their main cause of death. Since there are billions of cells in the human body, cancer can form anywhere. Human cells usually grow and divide to make new cells when the body needs them. Because of getting older or being hurt, old cells die and are replaced by new ones. Still, this systemic mechanism breaks down when cancer cells multiply. As cells become more abnormal, old or damaged cells live when they should die, and new cells form when they are not needed. These extra cells might keep making more and more of themselves, which could lead to the growth of tumors. There are two kinds of tumors: benign and cancerous. A benign tumor is a growth of cells that can't spread to other parts of the body. A malignant tumor, on the other hand, can spread. Metastasis is the process by which a disease moves to other parts of the body. Cancer comes in many forms, like colon, lung, and leukemia.

Since the beginning of the 1800s, lung cancer has become a lot more common. Lung cancer can be caused by many things, such as smoking, being around asbestos, being around radon gas, or being around someone who smokes. SCLC is less common and moves and gets worse more slowly than non-small cell lung cancer. SCLC makes tumors grow faster and bigger than smoking does, and it can spread fear throughout the body. They usually start in the bronchi, which are in the middle of the chest. The death rate from lung cancer is affected by the total number of cigarettes smoked. Using tobacco is a major cause of death and illness. Usually, lung cancer starts in the bronchial tree's walls or epithelium. But lung disease can affect any part of the respiratory system, and it can start anywhere. Most people with lung cancer are between the ages of 55 and 65, and the disease usually doesn't show up for many years. There are two main kinds of lung cancer. They are also called, respectively, cancer of the oat cell and lung cancer with tiny cells. Each subtype of lung cancer is treated differently and grows and spreads in its own way. "Mixed small cell/large cell cancer" is a term for a type of cancer that has traits of both small cell and large cell cancers. This document is put together as follows: The second part talks about work that is similar to this study, and the third part talks about the materials and methods used in this study. Then, in Section 4, results and evaluations of performance are given. The end of section five is the conclusion.

Related work

Krishnaiah, V., G. Narsimha et al. [1] proposed a method to predict Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques. In this study, Naive Bayes, ODANB, NCC2, Data Mining, Classification applied for train and test this system.

Yongqian Qiang et al. [2] developed a system to predict the Diagnostic Rules of Peripheral Lung Cancer Preliminary Study Based on Data Mining Technique.

Murat Karabhatak et al. [3] invented a system to create an expert system for detection of breast cancer based on association rules and neural network.

Y. Xie *et al. [4] developed a system to detect* Early lung cancer diagnostic biomarker discovery using machine learning. In this study, they collected total 110 lung cancer patients and 43 healthy individuals data to train and test their method. In this study, six machine learning techniques of K-nearest neighbor (KNN), Naïve Bayes, AdaBoost, Support Vector Machine (SVM), Random Forest, and Neural Network with 10-cross fold technique were used for the early lung tumor prediction based on the metabolomic biomarkers features.

D. M. Abdullah et al. [5] *proposed a system to detect* Lung cancer Prediction and Classification based on Correlation Selection method using machine learning techniques. They used UCI dataset to train and test their method. In this study, using SVM authors obtained 95.56% accuracy. While using KNN and CNN they obtained 88.40% and 92.11% accuracy, respectively.

C. H. Huang *et al. [7] proposed a system to detect Lung Cancer* Using a Chemical Sensor Array and a Machine Learning Technique. They used lung cancer cases and non-tumour controls between 2016 and 2018 for train and test their method. In this study, using SVM internal validation authors obtained 92.7% accuracy. While using LDA internal validation, LDA external validation and SVM external validation they obtained 90.2%, 85.4% and 85.4% accuracy, respectively.

J. Pati et al. [10] invented a system to detect Gene Expression Analysis for Early Lung Cancer Prediction using Machine Learning techniques and an Eco-Genomics approach. They used Kent Ridge Bio-Medical Dataset Repository. In this study, using SMO authors obtained 91.6667% accuracy. While using Multi-Layer Perceptron and Random Sub Space they obtained 86.6667% and 68.3333% accuracy, respectively.

Methodology

Data set:

Survey Lung Cancer is a familiar and commonly used data set for the prediction of Lung Cancer. (survey lung cancer | Kaggle). This data set consists of 310 rows and 16 columns. The attributes included in the column are gender, age, smocking, yellow_fingers, anxiety, peer_pressure, chronic disease, fatigue, allergy, wheezing, alcohol, coughing, shortness, shallowing, chest pain, lung_cancer.

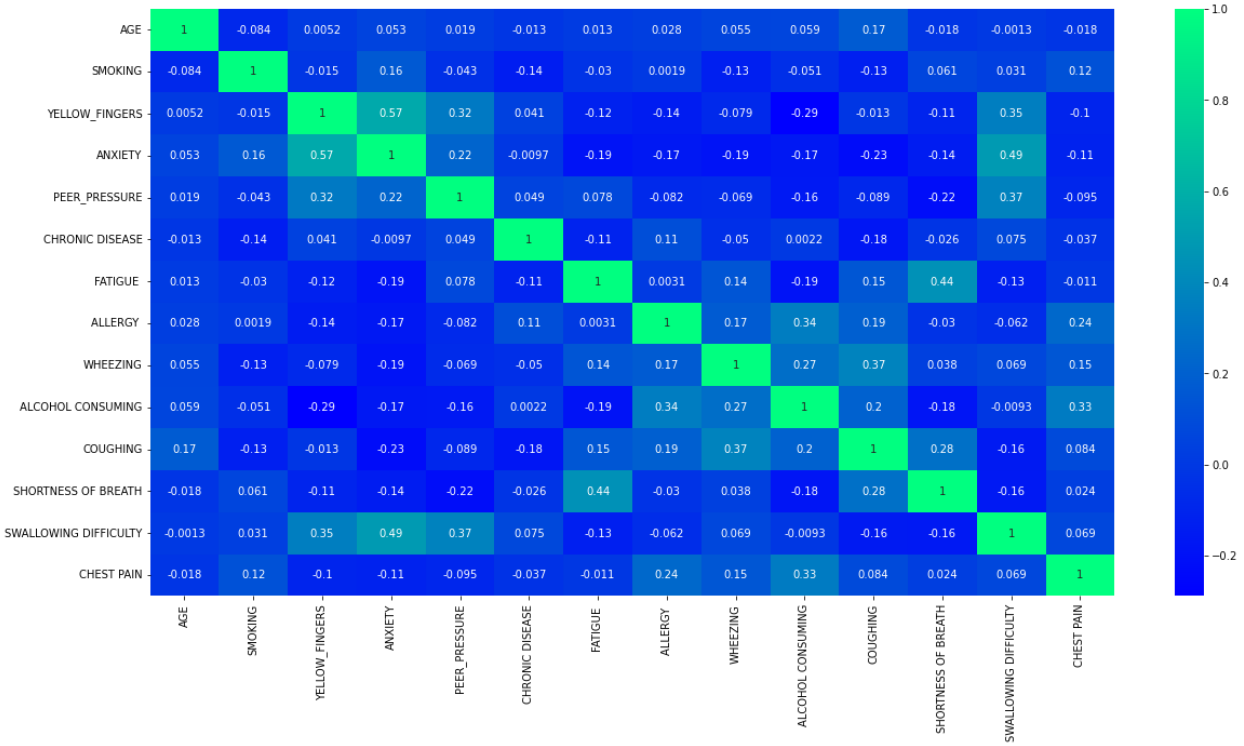


Figure 1: Heatmap for Lung Cancer dataset

Proposed method:

The research proposes a method for predicting and categorizing the many forms of lung cancer. The block diagram for the proposed work is depicted in Figure 2. The proposed approach begins with data collection followed by the application of machine learning techniques, model installation, and evaluation.

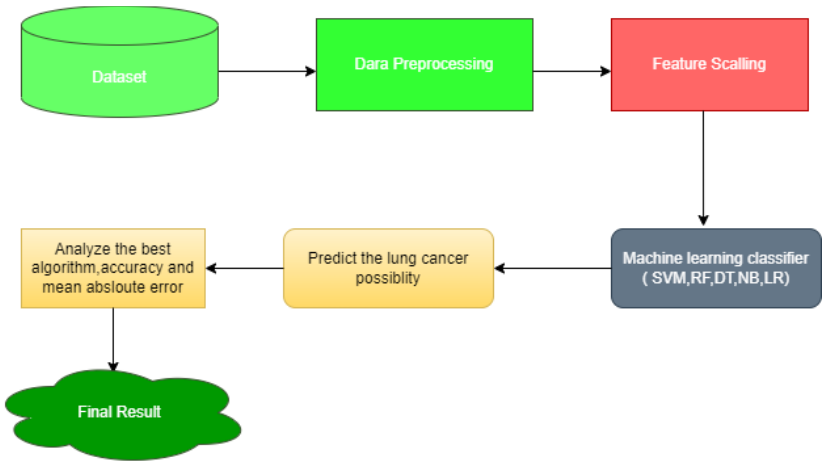


Figure 2: Block Diagram for Proposed method

Preprocessing:

a database consisting of hospitalized lung cancer patients that was created for the sole purpose of providing projections on incidences of lung cancer. Recently, researchers working in a wide range of disciplines have been drawing attention to the relevance of the way in which data are prepared [15]. The preparation of the data has made use of three key Python libraries, namely NumPy, Pandas, and Matplotlib. These libraries have been applied. Throughout the whole of the procedure for discovering and handling the missing data, a computing strategy that is centered on the mean has been applied. In the end, we used the Scikit learn libraries to import the train test split technique from the sklearn.model selection packages in order to split the dataset into a training section consisting of seventy-five percent of the data and a test portion consisting of twenty-five percent of the data. Because of this, we were able to proportionally divide the dataset into its many parts.

Machine learning algorithms:

Random Forest (RF):

The Random Forest technique is a kind of supervised learning that has applications in machine learning for solving problems involving both classification and regression. It is a sort of learning that is known as "ensemble learning," and it works by integrating the results of numerous classifiers in order to create predictions and enhance the model's overall performance. It includes a number of decision trees that may be applied to different subsets of the dataset. Determine the mean value in order to enhance the predictive

power of the model. A random forest should have between 64 and 128 trees in it. The greater the number of trees that are used, the more accurate the algorithm will be. Each tree in the algorithm casts a vote to determine how a new collection of data or object should be classified, and the algorithm then makes a prediction about the eventual outcome based on the result that received the most votes. The random forest method is a quick one that is able to handle incorrect and missing data in a competent manner.

Naive Bayas (NB):

A supervised learning method known as the Naive Bayes classifier is used to generate predictions about an item based on how likely it is to be that object. These predictions are based on the object's likelihood of being that object. The method is referred to as "Naive Bayes" due to the fact that it is based on Bayes's theorem and operates on the presumption that the variables do not interact with one another. The Bayes theorem relies on the idea of conditional probability, which may be defined as the likelihood of occurrence (A) in the case that occurrence (B) has already taken place. The following is the formula for Bayes's theorem:

$$P(X|Y) = \frac{P(Y|X)P(X)}{\mathrm{P(Y)}}$$

Finding a solution to an issue using the Naive Bayes classifier is one of the most effective methods to do it. A naïve Bayesian model is simple to construct and functions competently even when applied to a substantial quantity of data. Its primary function is to organize texts into categories.

Decision Tree (DT):

The majority of the time, decision trees are used to solve classification issues; however, they may also be used to tackle regression problems. A decision tree is an example of a supervised learning method. It is compatible with variables that remain constant as well as those whose values change over time. It illustrates a structure that is similar to a tree, complete with nodes and branches. It begins at the root node and progressively develops into further branches until it reaches the leaf node. The central node displays the properties of the dataset, the branches demonstrate the guidelines for decision-making, and the leaf nodes illustrate the answer to the issue.

Support Vector Machine (SVM):

An approach for supervised learning that is also capable of solving classification and regression issues is referred to as a support vector machine, which is also referred to by its acronym, SVM. However, it is most often used to issues involving the organization of items into groups. The purpose of support vector machines (SVM) is to create a "decision border" or "hyperplane" that categorizes datasets into several groups. The names of the data points that contribute to defining the hyperplane are referred to as "support vectors," which is also the name of the procedure. In the real world, SVM may be used to detect faces, categorize photos, discover novel medications, and a variety of other tasks.

Logistic Regression (LR) :

"Logistic regression" is the name of one of the most-used Machine Learning algorithms. It's a type of "supervised learning." It is used to predict the categorical dependent variable based on a set of independent variables.

Logistic regression is a way to figure out what will happen with a categorical dependent variable. The result must be a discrete or categorical value because of this. It can be Yes or No, 0 or 1, true or false, etc., but instead of giving the exact value of 0 or 1, it gives the probabilistic values that lie between 0 and 1. Logistic Regression is similar to Linear Regression, but it is used in a different way. Linear regression is used to solve regression problems, while logistic regression is used to solve classification problems. In logistic regression, instead of a regression line, we fit a "S"-shaped logistic function that predicts two maximum values (0 or 1). The curve from the logistic function shows how likely something is, like whether the cells are cancerous or not, whether a mouse is overweight or not based on its weight, and so on. Logistic Regression is an important machine learning algorithm because it can use both continuous and discrete datasets to give probabilities and group new data.

Function for Logistic Regression:

$$\text{Sig}(x) = \frac{1}{1 + e^{-x}}$$

Performance Evolution Matrices

Matrix of Confusion

The Confusion Matrix is a visual exam that may be used to evaluate one's level of deep learning. The forecast class results are shown in the columns of a Confusion Matrix, while the actual class results are displayed in the rows [54]. This matrix contains all of the raw data on the assumptions that a classification model makes for a particular batch of data. To determine how well a model works by testing it. It's a square grid, and each row displays the actual class that each instance belongs to, whereas each column displays the class that was anticipated for that instance. In a binary classification mission, the confusion matrix is a 22-by-22 table that displays the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

$$\begin{pmatrix} TP & FN \\ FP & TN \end{pmatrix}$$

The algorithms were evaluated using precision, recall, and the F-measure, all of which are metrics that are typically used in the communities of text mining and machine learning. True positive (TP) objects are those that have been correctly labeled as belonging to the class; false positive (FP) items are those that have been incorrectly labeled as belonging to a certain class; false negative (FN) items are those that have been incorrectly labeled as not belonging to a certain class; and true negative (TN) items are those that have been correctly labeled as not belonging to a certain class. There are four different types of classified items: true positive (TP), false positive (FP

The recall is calculated using the following formula [6][7], which takes into account the number of true positives as well as the number of false negatives:

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

The recall is also called the "absolute positive rate" or the "sensitivity." Precision, which is also called "positive predictive rate," is measured by the number of true positives and false positives:

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

F-measure is a way to measure both accuracy and recall. It is given as:

$$\mathrm{F} = \frac{(1+\beta 2)\times \mathbf{Recall} \times \mathbf{Precision}}{\beta 2 \times (\mathbf{Precision} + \mathbf{Recall})}$$

Experiments and Results:

This experiment is conducted by using lung cancer survey dataset which is comprises of 13 symptoms and 309 patients. The experiments were conducted on a local workstation using Jupyter Notebook, whose specs are supplied. Different python libraries (3.7.12) are used to experiment with NumPy (1.0.1), Pandas (1.11.0), Scikit-learn (1.0.1), and Matplotlib (3.5.1). (3.5.1).

Based on the experiment, we presented the findings in the outcome section. First of all this experiment preprocess the data and a mean calculation strategy has been adopted to discover and handle the missing data. For feature scaling, this research used a standardization method in which the independent variables in a dataset remain within a specific range. After preprocessing this PCA analysis is used for feature selection. Fit model is applied to modeling the dataset. According to the studies, six different projected result released because of six machine learning algorithm is employed which are random forest, decision tree, support vector machine (SVM), Gaussian Naïve Byes (NB), K means and Logistic Regression. This research mainly predicting the lung cancer of patients according to their symptoms. According to the experiment support

vector machine (SVM) outperformed the other five machine learning algorithms with the accuracy of 93 percent.

Machine Learning Algorithms	**Before Feature Selection**				**After Feature Selection**			
	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
Support vector machine (SVM)	82%	82%	100%	90%	93%	95%	96%	96%
Naïve Byes (NB)	83%	87%	94%	90%	91.59%	94%	97%	95%
Random Forest (RF)	91%	90%	97%	95%	90.91%	93%	96%	95%
Logistic Regression (LR)	83%	84%	98%	95%	91.61%	94%	96%	95%
Decision Tree (DT)	83%	89%	91%	90%	88.03%		94%	94%

Table 1 shows improvement of algorithms performance after PCA feature selection and 10 times k fold cross validation.

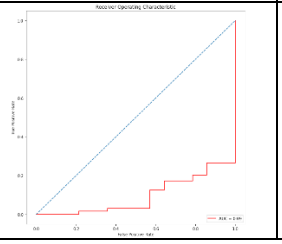
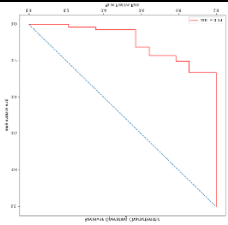
Support vector machine (SVM)	
Before Feature Selection	After Feature Selection
	

Table 2

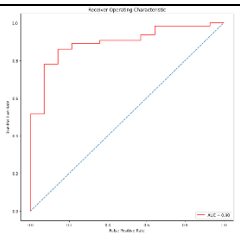
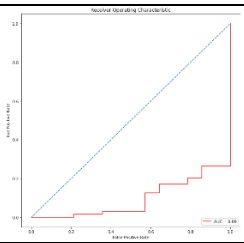
Naïve Byes (NB)	
Before Feature Selection	After Feature Selection
	

Table 3

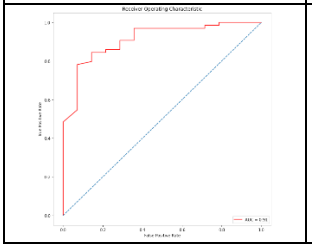
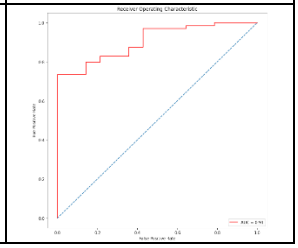
Random Forest (RF)	
Before Feature Selection	After Feature Selection
	

Table 4

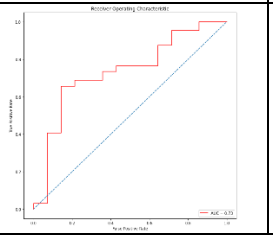
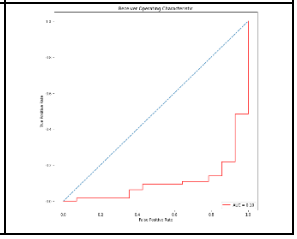
Logistic Regression (LR)	
Before Feature Selection	After Feature Selection
	

Table 5

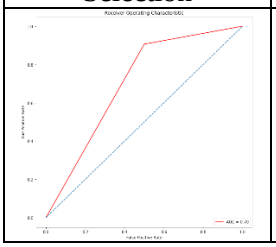
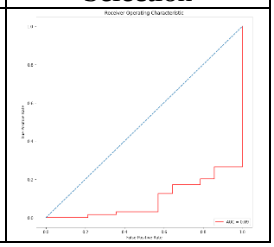
Decision Tree (DT)	
Before Feature Selection	After Feature Selection
	

Table 6

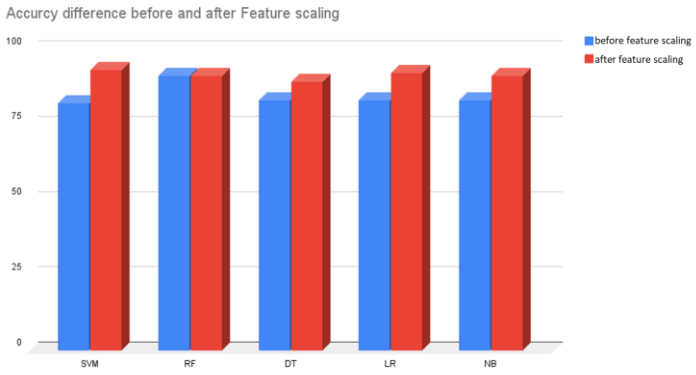


Figure 3: After and Before feature optimization accuracy chart

The following table (2-6) provides a comparison of the ROC curve for several kinds of machine learning methods. This is an attempt to demonstrate the feature selection process of machine learning algorithms before and after it has been applied. The SVM method produces the best results in this experiment.

Discussion

According to the findings of this research, the suggested method is capable of accurately diagnosing lung cancer in its early stages when combined with an approach based on machine learning and an enhanced technology for analyzing data. In this work, comprehensive validation and correlation checks were performed on the dataset. Table 1 compares the performance of a number of different machine learning algorithms based on a number of different metrics, including standard deviation, ROC, AUC, accuracy, precision, recall, and f-1 score. Following extensive investigation, we got to the conclusion that none of the classifiers come anything near to having an accuracy of one hundred percent. The SVM performed very well in comparison to the other classifiers. In addition, the performance metrics of the different machine algorithms are projected in Table 1-7 and Figure 1-3 respectively, demonstrating how effective these algorithms are.

References

[1] Krishnaiah, V., G. Narsimha, and Dr N. Subhash Chandra. "Diagnosis of lung cancer prediction system using data mining classification techniques." International Journal of Computer Science and Information Technologies 4.1 (2013): 39-45.

[2] Yongqian Qiang, Youmin Guo, Xue Li, Qiuping Wang, Hao Chen, & Duwu Cuic 2007 .The Diagnostic Rules of Peripheral Lung cancer Preliminary study based on Data Mining Technique. Journal of Nanjing Medical University, 21(3):190-195

[3] Murat Karabhatak, M.Cevdet Ince 2008. Expert system for detection of breast cancer based on association rules and neural network. Journal: Expert systems with Applications.

[4] Y. Xie *et al.*, "Early lung cancer diagnostic biomarker discovery by machine learning methods," *Translational Oncology*, vol. 14, no. 1, Jan. 2021, doi: 10.1016/j.tranon.2020.100907.

[5] D. M. Abdullah, "Lung cancer Prediction and Classification based on Correlation Selection method Using Machine Learning Techniques", doi: 10.48161/Issn.2709-8206.

[6] Sugianela, Y., & Ahmad, T. (2020, February). Pearson Correlation Attribute Evaluation-based Feature Selection for Intrusion Detection System. In 2020 International Conference on Smart Technology and Applications (ICoSTA) (pp. 1-5). IEEE.

[7] Demisse, G. B., Tadesse, T., & Bayissa, Y. (2017). Data mining attribute selection approach for drought modeling: A case study for Greater Horn of Africa. arXiv preprint arXiv:1708.05072.

[8] G. A. P. Singh and P. K. Gupta, "Performance analysis of various machine learning-based approaches for detection and classification of lung cancer in humans," *Neural Computing and Applications*, vol. 31, no. 10, pp. 6863–6877, Oct. 2019, doi: 10.1007/s00521-018-3518-x.

[9] Dayananda Sagar College of Engineering, Institute of Electrical and Electronics Engineers. Bangalore Section, and Institute of Electrical and Electronics Engineers, *2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA 2020) : conference proceedings : 5-7 March, 2020*.

[7] C. H. Huang *et al.*, "A study of diagnostic accuracy using a chemical sensor array and a machine learning technique to detect lung cancer," *Sensors (Switzerland)*, vol. 18, no. 9, Sep. 2018, doi: 10.3390/s18092845.

[8] N. Banerjee, "Prediction Lung Cancer-In Machine Learning Perspective," 2020.

[9] M. Imran Faisal, S. Bashir, Z. Sikandar Khan, and F. Hassan Khan, "An Evaluation of Machine Learning Classifiers and Ensembles for Early Stage Prediction of Lung Cancer."

[11] "AN EXTENSIVE REVIEW ON LUNG CANCER DETECTION USING MACHINE LEARNING TECHNIQUES," *Journal of critical reviews*, vol. 7, no. 14, Jul. 2020, doi: 10.31838/jcr.07.14.68.

[12] Q. Gu *et al.*, "Machine learning-based radiomics strategy for prediction of cell proliferation in non-small cell lung cancer," *European Journal of Radiology*, vol. 118, pp. 32–37, Sep. 2019, doi: 10.1016/j.ejrad.2019.06.025.

[13] J. Pati, "Gene expression analysis for early lung cancer prediction using machine learning techniques: An eco-genomics approach," *IEEE Access*, vol. 7, pp. 4232–4238, 2019, doi: 10.1109/ACCESS.2018.2886604.

[14] J. M. Luna *et al.*, "Predicting radiation pneumonitis in locally advanced stage II–III non-small cell lung cancer using machine learning," *Radiotherapy and Oncology*, vol. 133, pp. 106–112, Apr. 2019, doi: 10.1016/j.radonc.2019.01.003.

[15] **Prediction of dengue incidents using hospitalized patients, metrological and socio-economic data in Bangladesh: A machine learning approach**
Dey SK, Rahman MM, Howlader A, Siddiqi UR, Uddin KMM, et al. (2022) Prediction of dengue incidents using hospitalized patients, metrological and socio-economic data in Bangladesh: A machine learning approach. PLOS ONE 17(7): e0270933. https://doi.org/10.1371/journal.pone.0270933