Lung Cancer Prediction using Machine Leaning Techniques

|  |  |  |
| --- | --- | --- |
| Aadi Poddar School of Computer Engineering KIIT Deemed UniversityBhubaneswar, Odisha, INDIA aadipoddarmail@gmail.com | Mahendra Kumar Gourisaria *School of Computer Engineering KIIT Deemed to be University* Bhubaneshwar, Odisha, India mkgourisaria2010@gmail.com | Monika Padhi  Viswas School and College of Nursing Bhubaneswar, Odisha, India  monikapadhi444@gmail.com |
| Anjishnu Saw  *School of Computer Engineering*  *KIIT Deemed to be University*  Bhubaneshwar, Odisha, India  23051895@kiit.ac.in | Saurabh Bilgaiyan *School of Computer Engineering KIIT Deemed to be University* Bhubaneshwar, Odisha, India saurabh.bilgaiyanfcs@kiit.ac.in | Tanvir Habib Sardar  *CSE, School of Engineering*  *Dayanand Sagar University*  Bangalore, Karnataka, India  tanvir-cse@dsu.edu.in |

Abstract – The most commonly diagnosed cancers in the world are Lung cancer. Beforehand prediction, detection and diagnosis of lung cancer has become essential. This paper draws light to the lung cancer predictions and detection using common life habits and attributes. Predictions are based on different parameters like gender and age of the person and their habits like smoking and alcohol consumption. The effectiveness of lung cancer prediction system helps the doctor and patients to know their lung cancer risk with minimal cost. The prediction helps people to take appropriate actions based on their lung cancer risk status. This paper involves various machine learning regression models like Lasso Regression, Linear Regression, Ridge Regression, Decision Tree and Random Forests. In this paper all the above models are compared, and it is found that Random Forest Regressor outperforms the best results in MSE, RMSE and R2 Score as compared to other models with the value 0.0228, 0.1536 and 0.2447.

Keywords—Lung Cancer Detection, Machine Learning, Predictive Modeling, Health data analytics, Regression Analysis

# Introduction

Lung cancer is responsible for increasing cancer deaths over the years, globally. It has been observed from the data 1.8 million new cases are reported every year. Lung cancer accounts for the highest mortality rate among both men and women. Smoking is the leading cause of lung cancer and cancer deaths; it leads to 85% of all cases [1]. Most people are diagnosed at an advanced stage of lung cancer which results in a low survival rate. In the past few years 27% of all cancer deaths and 14% of all caner diagnoses were due to Lung Cancer [2]. To solve this problem, machine learning engineers and data scientists are working to predict lung cancer for the people so that lung cancer can be predicted and diagnosed as early as possible. Predicting and determining cancer in an early stage is better and has a higher chance of cure and the patient has a higher rate of survival. The various aspects of Machine Learning and Data is cleaned and sorted before training the models removing unnecessary data attributes which do not contribute in the classification and prediction process.

The machine learning models used in this paper use attributes regarding basic information about the person like gender and age. It consists of other attributes like habits of the person like smoking and alcohol consumption. It takes more information about how the person feels in day-to-day life like Fatigue, feels exhausted, feels wheezing in which they make a high-pitched sound while breathing, person has shortness of Breath, Swallowing Difficulty, Chest pain, yellow fingers and is under Peer Pressure. All these variables and attributes are combined to give the lung cancer result which is the target variable.

Lung cancer prediction using Machine Learning Regression models is getting popular day by day helping doctors catch the diseases earlier and save life. It is highly important to estimate and predict the risk of cancer accurately to the patients and doctors for fulfilling proper medication and ensuring timely and accurate treatment.

Detection and diagnosing of lung cancer is a great challenge encountered by doctors and researchers now a days. Machine Learning models approaches recognize the main attributes of complex datasets of lung cancer. A CAD (Computer-Aided Diagnosis) emerged in the early 1980s to improve and enhance the survival rate and efficiency that helped doctors globally in interpreting medical images [3].

Lung cancer prediction using the basic human habits are necessary because existing clinical techniques such as X-Ray and other imaging procedures for detecting cancer require complex hardware. These techniques are expensive, and this raises the need for machine learning models that are comparatively cheaper and easy to access. Simple and easy to use machine learning models are essential for the detection of cancer in remote locations and places where medical facilities have not been established yet. It is essential for a vast variety of people who are not permitted for X-Rays and other surgical tests that require radiation entering their body [4].

ML algorithms can also be used in different other fields like maize leaf disease diagnosis [5], rice disease prediction [6], diabetes prediction [7] etc. The rest of the paper is organized like: Section II. Literature Review, covering significant publications on the subject; Section III. tells about Methods and Materials, providing details on the classifiers and strategies employed; Section IV covers Implementation and Results, contrasting the performance of various regression models to determine the most effective model; and Section V. Conclusion and Future Work, summarizing findings and proposing future research directions followed by references followed by reference

# Related Work

This section provides the review of research various works performed in the field of Lung Cancer Detection. Chinmayi Thallam (2020) [8] examined the performance on the dataset from The Lung Image Database Consortium (LIDC-IDRI). The dataset was trained on a number of algorithms including regression and classifiers-Random Forest and Voting Classifier. The most accurate results were obtained using Voting Classifier of 99.5%. The study showed how K-NN holds good only with balanced data in the dataset and is sensitive to outliers. It makes it difficult to show the problem to the artificial neural network when it works with numerical data and the network duration is unknown. In the study, Muhammad Imran Faisal (2018) [9] examined the performance from the dataset from UCI repository. The dataset was trained on various models like regression and classifiers- Random Forest and Decision Tree. The best accuracy was 85.75% and was achieved through Gradient Boosted Tree. The Gradient Boosted Tree excelled among all the classifier models.

In the study of Lung Cancer Risk Prediction with Machine Learning Models of E. Dritsas (2022) [10] the research work makes use of supervised learning techniques for the process of identification of patients with lung cancer based on several attributes and features. The study uses many machine learning models like Rep Tree, RotF and AdaBoost which were reviewed based on their accuracy and precision scores. The RotF model outperformed all other models with 99.3%. accuracy.

In the study of prediction of Lung cancer using various ensemble models by Muntasir Mamun (2022) [11] the study took into consideration some previous researches and studies related to lung cancer prediction machine learning models and compared the performances to the models. Models such as Light GBM, XG Boost, bagging and AdaBoost were developed to predict lung cancer using the lung cancer dataset. According to the overall analysis, XG Boost outperformed all the models with an accuracy of 94.42%.

The detection of lung cancer for symptom-based diagnosis of lung cancer was studied by K. Sivanagireddy (2023) [12]. Several machine learning regression models like Linear, Logistic and Logistic regressions were trained on the dataset to accomplish lung cancer detection. This study calculates the R squared values based on attributes such as long-term illness to diagnose lung cancer. The multiple regression outperforms other regression models as with a higher accuracy of 96% in predicting lung cancer.

The detection of lung cancer by CT scan pictures and histopathology images was studied by B. Mamatha (2023) [13] for the identification of lymph node involvement on histopathology slides. The results demonstrate that detection accuracy improves when histopathological tissues are analyzed. Several algorithms, including CNN's quasi-convex GD, VGG19, and Resnet50, were used for the prediction analysis.  The CNN GD model gave the best results, performing at 99.84% accuracy. It outperformed other models that were tested with a F-Score of 98.25%.

The detection of lung cancer considering the various symptoms like age, gender, shortness of breath, and chest pain, alcohol consumption, chronic disease, swallowing difficulty, anxiety and peer pressure was done by J. Surendiran (2022) [14]. The dataset has been collected from Kaggle, the data is about patients who are affected by lung cancer and the symptoms they have. Various machine learning algorithm like linear regression, logistic regression, logarithmic regression, multiple regression and exponential regression have been used to predict lung cancer in this paper. It was found that multiple regression algorithm outperforms other models as the best performing regressor model achieving an accuracy of 96% in predicting lung cancer from the symptoms that have been determined in the patients.

A constructive and detailed analysis of all the literature reviews alongside their performance score i.e., accuracy is shown in Table.1

Table I. Analysis of Literature Review

|  |  |  |  |
| --- | --- | --- | --- |
| Year | Model | Dataset | Accuracy |
| 2020 [8] | Support Vector Machine | The Lung Image Database Consortium (LIDC-IDRI) | 95% |
| 2018 [9] | Gradient Boosted Tree | UCI repository | 85.71% |
| 2022 [10] | RotF | UCI repository | 99.3% |
| 2022 [11] | XG Boost | MIMIC-III | 94.42% |
| 2023 [12] | Multiple Regression | Online Dataset from Kaggle | 96% |
| 2023 [13] | CNN GD model | CT scans from the Cancer Imaging Archive | 99.84% |
| 2022 [14] | Multiple Regression | Online Dataset from Kaggle | 96% |

# PROPOSED METHODS

Different machine learning models work differently and it is entirely up to us as to which model we should train to obtain the best results. In this paper we have applied different regression techniques which includes algorithms like Linear Regression, Ridge Regression, Lasso Regression, Random Forest Regression and Decision Tree Regression. These regression algorithms are discussed further, along with the results obtained after training the models. The subsection A. described the Dataset Used, B. Hardware and Software used, C. Preprocess Data, D. talk about Splitting the Data, E. Model Workflow Diagram, F. Correlation Matrix, G. Model Used, H. Evaluation Metrices.

## Dataset Used

The dataset has been derived from Kaggle [15] which was created and maintained by Atharva Khairnar. The algorithms employed for predictions are based on real-world data, which is often difficult to handle, unbalanced, and incomplete. The dataset consists of 309 records and 16 attributes. Attributes include basic demographic information (Gender, Age), health-related symptoms (YELLOW\_FINGERS, ANXIETY, FATIGUE, WHEEZING, COUGHING, CHEST PAIN), habits (ALCOHOL CONSUMPTION, SMOKING), and additional factors (ALLERGY, PEER\_PRESSURE). The target attribute, LUNG\_CANCER, indicates the presence or absence of lung cancer.

## Hardware and software

The training and testing of all models were done on a workstation running Windows 11 Home Single Language OS and with an Intel 12th Gen Intel(R) Core (TM) i7-1255U 1700 MHz CPU, 10 cores, 12 logical processors and 16 GB RAM. Scikit-Learn modules for the training of the models used Python 3 carried out on Google Collab.

## Preprocess Data

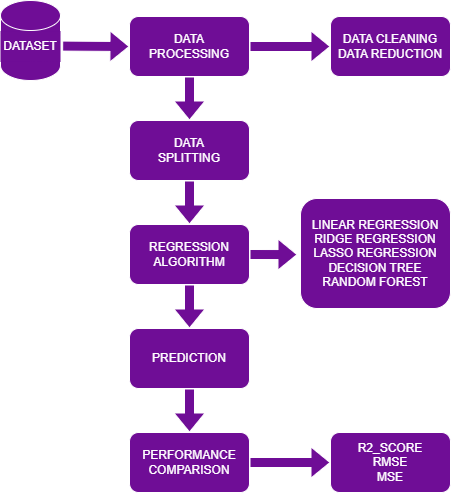
The Gender attribute, containing values M (male) and F (female), is replaced with numerical values. Categorical data is converted to binary by mapping M to 1 and F to 0. The target variable (LUNG\_CANCER) is encoded numerically, where 1 represents "YES" and 0 represents "NO." Binary encoding is ensured by replacing values such as 2 (originally denoting "YES" in features like smoking or anxiety) with 1, standardizing all features to a 0/1 binary format.

## Data Splitting

Input features (e.g., smoking, age, symptoms) are stored in one variable, while the target variable (LUNG\_CANCER) is stored separately. The dataset is split into train and test data sets, with 20% reserved for testing and 80% for training. The training dataset set is used to train the machine learning model. The testing set of the regression models evaluates performance on unseen data. Dependent variables are predicted using the trained algorithm, and results are compared against original values to minimize error.

## Model Workflow Diagram

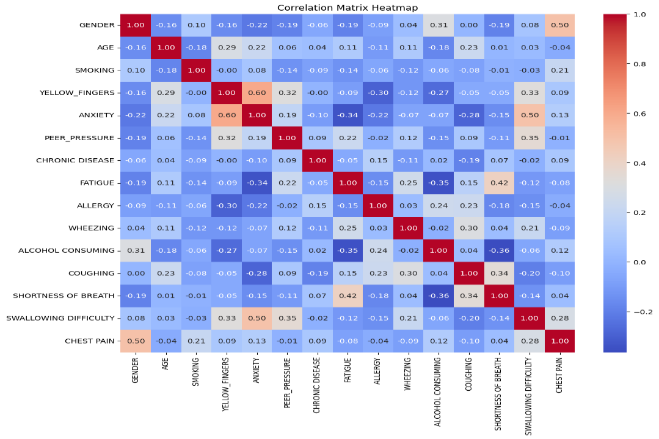
A well-designed workflow should be framed from the dataset to finding the scores and results of the algorithms in order to make a promising machine learning model. The flowchart of all the necessary processes required to develop a model and test it on a processed dataset can be seen in Fig-1.



1. Dataset Flow Chart

## Correlation Matrix

The correlation matrix is used to see the correlation between the attributes. Highly correlated attributes and the target attribute are selected. It is the table to display the correlation coefficients between the attributes and variables of the dataset. It is useful for visualizing the relationships between the variables. A perfect positive correlation is represented by a correlation coefficient of 1, a perfect negative correlation by a correlation value of -1, and no correlation is represented by a correlation coefficient of 0 (usually Pearson's). The correlation matrix of all the relevant attributes of the dataset post data processing has been plotted in Fig-2.



1. Correlation Matrix of all features after processing and cleaning the dataset

## Technology Models used

We have used Linear Regression, Ridge Regression Lasso Regression, Decision Tree and Random Forest Regression. A brief description of all the models are given below.

### Linear Regression Model

This model is used between the input attributes and the target variable that finds a linear relationship between the attributes [16]. A model is created with the training dataset which is a hypothesis that takes the new input attribute and gives the output of the target attribute. Performance metrices are used to verify whether this model is performing well. A model is created using the training dataset, forming a hypothesis that predicts the target variable from input features. Performance metrics are employed to validate the model. In Linear Regression, the best-fit line is determined to model the relationship between input features (x-axis) and the target variable (y-axis), this has been shown in Eq-1.

|  |  |  |
| --- | --- | --- |
|  |  | (1) |

*β₀* is the Intercept. When *x = 0* then *y = β₀*. This shows that at what point the line is meeting the y axis. *β₁, β₂, …, βₙ* are the slope or coefficients for the input features *x₁, x₂, …, xₙ*

*ϵ* denotes the error term that the unnecessary noise in the relationship between attributes. Algorithm is best to use when the independent and dependent variables have linear relationship, but it is always assumed that there is a straight-line relationship between the dependent and independent attributes. Linear Regression is optimized for fast speeds.

### Ridge Regression

This is applied to prevent overfitting. It is used in such a way that the model should never overfit [17]. It introduces an L2 regularization term to the linear regression to ensure model stability, this has been shown in Eq-2.

|  |  |  |
| --- | --- | --- |
|  |  | (2) |

(lambda) is the hyperparameter controlling the penalty strength and are the coefficients.

Coefficients are shrunken towards zero. On changing the values, a best fit line is obtained. After the line a slop is obtained which will decrease each time. After sometime this value will not decrease it will be a minimal and smaller value. Iterations are specified that is how many times the data is trained. Iteration is also a hyperparameter. Obtaining a 0 is not possible because then it will be an overfitting model. It is best to use for datasets where overfitting is a risk.

### Lasso Regression

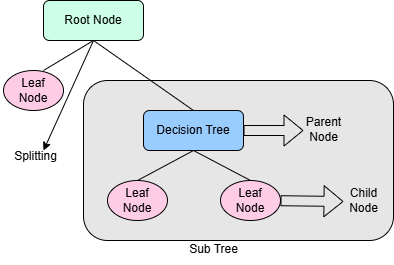
This model adds the sum of mod slope instead of square of slope. This Mod of slop work is to do feature selection [18]. On summing the values, the slope value will be extremely small for the features that are not playing a significant role. The entire feature is then neglected. Eq-3 mathematically shows the formula of Lasso Regression.

|  |  |  |
| --- | --- | --- |
|  |  | (3) |

Ridge Regression does square of the slope which increases that particular value but because of the mod in Lasso Regression that value will not increase. In summary this neglects all features that are not at all important for this model. The major benefits of this model are to prevent overfitting and in cases where there are many features where many of them are not important. It performs feature selection. λ (lambda) is a hyperparameter which will be found out be cross validation. The cost function is to be reduced in such a way that it will never become zero but will be reduced based on the λ and the slope value.

### Decision Tree

This is a sort algorithm where data is split according to a certain parameter [19]. Feature are selected by Information gain. Decision Tree does not have a good time complexity for execution purpose. It splits all the features to its complete depth for training data, the accuracy is high and for test data the accuracy is low. It leads to low bias and high variance. This can be solved by pruning, but pruning is an extensive task for large datasets. Decision Tree Regression also uses Entropy, Information Gain, Gini Impurity to calculate the further results. Fig-3 shows the functioning of the Decision Tree regressor in detail with proper explanation of the root and leaf nodes.

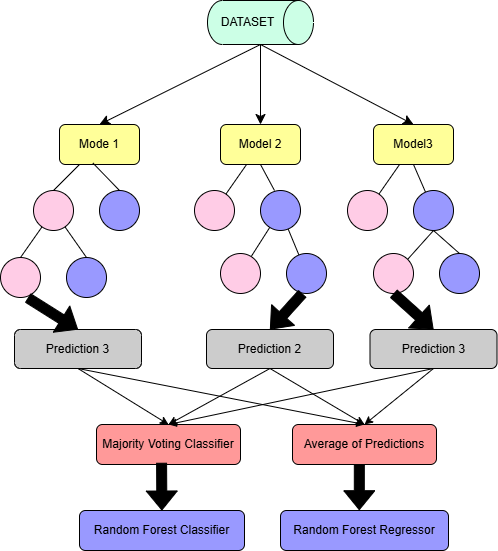


1. Decision Tree Regression

### Random Forest

The main problem with Decision Tree is that it leads to overfitting when it is created without a hyperparameter [20]. The main aim of Random Forest is to convert the high variance of decision tree to low variance. Random Forest is a bagging technique in which a dataset has multiple models. Each model is a Decision Tree that is going to be used in Random Forest. Each Decision Tree model creates Low bias and High Variance. Combining each Decision Tree model in the form of bootstrap aggregator the High Variance will be converted to low variance. This is because majority of the voting will be taken from these decision trees.

This is done by Majority Voting Classifiers. In Random Forest each model is created by Row Sampling and Feature Sampling. This means that some set of rows and some sets of features are given to each model. This leads to features getting repeated, records and datapoints also get repeated in the models. Each model will be an expert in predicting the specific data on which it was trained. Fig-4 shows the functioning of the Random Forest Regressor in detail with proper explanation taking different decision tree models.



1. Random Forest Classifier

## Evaluation Matrix

Machine Learning is based on the analysis and testing of dataset. On processing the model with certain algorithms, gives the predictions. The accuracy of the regression model is calculated which is how well the model predicted the results. A comparison was illustrated between the predicted results and the actual values to test the accuracy and behaviour of the model. Errors represent any false predictions made by the model. The evaluation metrics that are used are –

### Score

R square is the coefficient of determination, which ranges from 0 to 1. It is represented by a percentage. Higher value of R square represents higher accuracy of the prediction model. Eq-6 shows the derivation of R-Square.

|  |  |  |
| --- | --- | --- |
|  |  | (6) |

states the mean value of x

### Mean Squared Error (MSE)

Mean Squared Error (MSE) is the mean of the squared errors. It represents the difference between the mean squared of the predicted and the actual values. Eq-4 shows the derivation of MSE.

|  |  |  |
| --- | --- | --- |
|  |  | (4) |

in Eq-4 shows the number of predictions,

*x* represents the actual values

are the values predicted by the model

### Root Mean Square Error (RMSE)

Root Mean Square Error is the square root of the mean of the squared errors. It is the square root of Mean Square Error. The difference between the predicted and actual value is the error. Eq-5 shows the derivation of RMSE.

|  |  |  |
| --- | --- | --- |
|  |  | (5) |

# Results And Discussion

Machine learning models were evaluated on the dataset based on their performance to observe the robustness and effectiveness of the model. It can be evaluated in several ways such as by checking the Mean Squared Error (MSE), the coefficient of determination (R2\_Score) and the Root Mean Squared Error (RMSE). The Random Forest Regression outperforms other regression model by giving the best results in R2 Score. Table 2 shows the performance of the all the regression models used in this study, along with their performance scores.

Table II. Analysis of related work

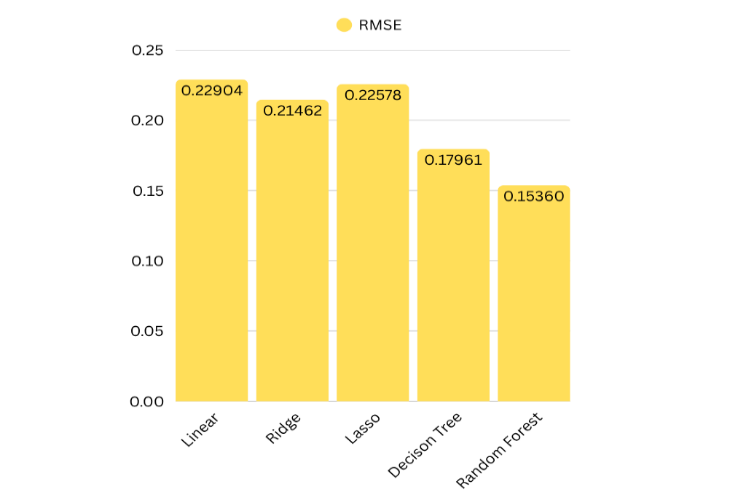
|  |  |  |  |
| --- | --- | --- | --- |
| Models | R2 Score | MSE | RMSE |
| Ridge Regression | -0.4755 | 0.0461 | 0.2146 |
| Linear Regression | -0.6804 | 0.0525 | 0.2290 |
| Lasso Regression | -0.6330 | 0.0510 | 0.2258 |
| Decision Tree Regressor | -0.0333 | 0.0323 | 0.1796 |
| Random Forest Regressor | 0.2447 | 0.0228 | 0.1536 |

It is clear that, the higher the score of the coefficient of determination (R-squared), lower is the Mean Squared Error (MSE) and the Root Mean Squared Error (RMSE). These value and interpretation help to predict which model performs best according to the given dataset giving the best possible results accurately. Fig. 5 (a) and 5 (b) shows the graphical representation of R2 and RMSE.

Fig. 6 depicts the performance of various algorithms graphically in terms of MSE. It has been found out that Random Forest Regression outperforms other regression models with a 0.2447 R-Square score, MSE score of 0.0228 and 0.1536 RMSE score. The least performing regression



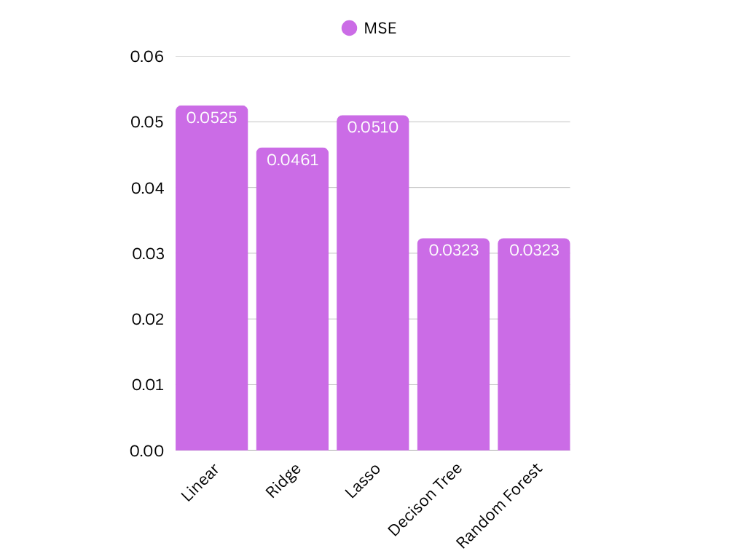
(a)



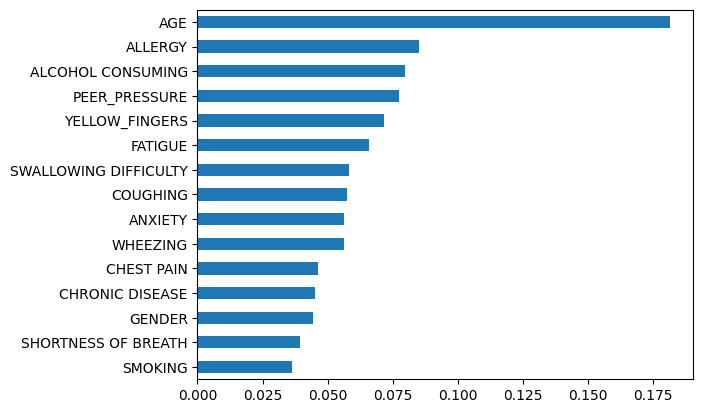
(b)

1. Measure of Performance of various algorithms for (a) R\_2 Score (b) RMSE.

model was found to be Linear Regression with a -0.6804 R-Square score, 0.0525 MSE score and a 0.2290 RMSE score. Age comes as the feature of most importance in the Random Forest training model. Fig. 7 graphically depicts all the features and attributes of the dataset with their importance in the random forest training model.



1. Measure of Performance of various algorithms for MSE.



1. Importance of Attributes of the Dataset

# Conclusion And Future Work

This paper uses the basic life habits and information about the person to predict lung cancer. Many evaluation techniques like Root Mean Square Error are used on different machine learning models. Regression Models used to predict cancer are Linear Regression, Ridge Regression and Lasso Regression. Classifier Models like Decision Trees Classifier, Logistic Regression, and Random Forests were proved to have more accuracy in predicting lung cancer. This helps the patients and doctors to detect lung cancer easily just by their habits and basic information. This paper shows that Random Forest Regression model performs best among as compared to other models. It outperforms as the best regression model giving the best R2 Score, MSE and RMSE when trained and tested on the given dataset of lung cancer derived from the symptoms several patients.

For further research, the researchers can develop a mobile application to take the information and habits of the user on a daily basis and warn them if they are at a risk of lung cancer. This can be done by machine learning algorithms for predicting cancer more accurately. We can try our work on more bigger dataset.

##### REFERENCES

1. World Health Organization (WHO) Report about Lung Cancer, [Who Report About Lung Cancer](https://www.who.int/news-room/fact-sheets/detail/lung-cancer), 2023. <https://www.who.int/news-room/fact-sheets/detail/lung-cancer>
2. World Cancer Research Fund report about Lung Cancer 2022. <https://www.wcrf.org/preventing-cancer/cancer-statistics/lung-cancer-statistics/>
3. S. S. Raoof, M. A. Jabbar and S. A. Fathima. "Lung Cancer prediction using machine learning: A comprehensive approach." In *2020 2nd International conference on innovative mechanisms for industry applications (ICIMIA)*, pp. 108-115. IEEE, 2020
4. A. Worawate, A. Thirach, S. Marukatat, and T. Wilaiprasitporn. "Automatic lung cancer prediction from chest X-ray images using the deep learning approach." In *2018 11th biomedical engineering international conference (BMEiCON)*, pp. 1-5. IEEE, 2018.
5. P. Theerthagiri, A. U. Ruby, J. G. C. Chandran, T. H. Sardar, & A. Shafeeq BM, Deep SqueezeNet learning model for diagnosis and prediction of maize leaf diseases. Journal of Big Data, 11(1), 112, 2024.
6. S. Das, T. H. Sardar, & D. S. Sahana, An ensemble technique using genetic algorithm and deep learning for the prediction of rice diseases. In Machine Learning Hybridization and Optimization for Intelligent Applications, pp. 289-303, CRC Press, 2025.
7. A. Raj, J. J. Jena, D. Ghosh, S. Jain, T. H.Sardar, & M. K.Gourisaria, Early-Stage Diabetes Risk Prediction Using Machine Learning Techniques. In 2025 3rd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT), pp. 2289-2295, 2025
8. C. Thallam, A. Peruboyina, S. S. T. Raju and N. Sampath. "Early stage lung cancer prediction using various machine learning techniques." In *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pp. 1285-1292. IEEE, 2020.
9. M. I.Faisal, S.Bashir, Z. S.Khan, & F. H.Khan, An evaluation of machine learning classifiers and ensembles for early stage prediction of lung cancer. In 2018 3rd international conference on emerging trends in engineering, sciences and technology (ICEEST), pp. 1-4, 2018
10. E. Dritsas and M. Trigka, Lung cancer risk prediction with machine learning models. *Big Data and Cognitive Computing*, *6*(4), p.139, 2022
11. M. Mamun, A. Farjana, M. Al Mamun and M. S. Ahammed, "Lung cancer prediction model using ensemble learning techniques and a systematic review analysis." In *2022 IEEE World AI IoT Congress (AIIoT)*, pp. 187-193. IEEE, 2022.
12. K. Sivanagireddy, S. Yerram, S. S. N. Kowsalya, S. S. Sivasankari, J. Surendiran and R. G. Vidhya, "Early lung cancer prediction using correlation and regression." In *2022 International Conference on Computer, Power and Communications (ICCPC)*, pp. 24-28. IEEE, 2022.
13. B. Mamatha, D. Rashmi, S. K. Tiwari, A. P. Sikrant, A. Arokiaraj Jovith, and P. C. S. Reddy. "Lung Cancer Prediction from CT Images and using Deep Learning Techniques." In *2023 Second International Conference on Trends in Electrical, Electronics, and Computer Engineering (TEECCON)*, pp. 263-267. IEEE, 2023.
14. J. Surendiran, K. D. Kumar, T. Sathiya, S. S. Sivasankari, R. G. Vidhya, and N. Balaji. "Prediction of lung cancer at early stage using correlation analysis and regression modelling." In *2022 Fourth International Conference on Cognitive Computing and Information Processing (CCIP)*, pp. 1-6. IEEE, 2022.
15. A. Khairnar**, “**Lung Cancer Dataset” Kaggle – Does Smoking cause Lung Cancer (Version 1) 2021, Retrieved on 24th March 2025, <https://www.kaggle.com/code/khairnaratharva/addbost-98-387-knn-96-7741-lr-svm-dt-rf>
16. J. Gareth, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor. "Linear regression." In  “An introduction to statistical learning: With applications in python”, pp. 69-134. Cham: Springer international publishing, 2023.
17. G.C. McDonald, Ridge regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, *1*(1), pp.93-100, 2009
18. J. Ranstam, & J. A. Cook, LASSO regression. Journal of British Surgery, 105(10), 1348-1348, 2018.
19. D. B. Ville, Decision trees. Wiley Interdisciplinary Reviews: Computational Statistics, 5(6), 448-455, 2013.
20. S. J. Rigatti. Random forest. Journal of Insurance Medicine, 47(1), 31-39, 2017.