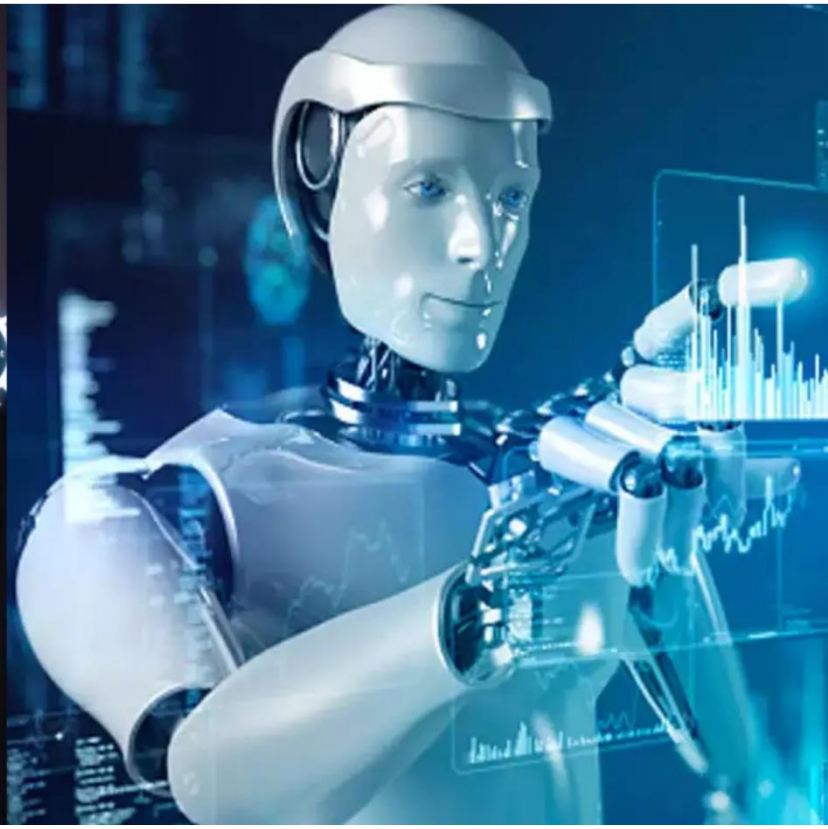
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 5, May 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405440

Lung Cancer Prediction through Regression Techniques

N Dhivya, Shalini S

Assistant Professor, Vivekanandha Institute of Information and Management Studies, Tiruchengode, Tamil Nadu, India

Master of Computer Application, Vivekanandha Institute of Information and Management Studies, Tiruchengode, Tamil Nadu, India

ABSTRACT: Nowadays, the changes in food habits and working culture among the individuals causing many health issues, in which cancer is most important life threatening disease amongst people. Lung cancer is most common among the cancer patients all over the world. Early detection and diagnosis of cancer helps in long survival for patients, whereas failure in early detection may cause fatal end. Data mining techniques like classification, regression and clustering are more powerful in disease detection. In this work, we proposed regression analysis on lung cancer dataset from UCI respiratory data. The regression algorithm used in this study are Lasso, Linear, Logistic and random forest regression. From our experimental results, it is observed that random forest regression achieves good accuracy.

KEYWORDS: : Smart Administration, Centralized System, Resource Management, Centralized Resource Dashboard, Web-Based Administration Panel ,Online Resource Management System

I. INTRODUCTION

One of the main reasons of non-unintentional death is cancer. From many surveys, it has been proven that lung most cancers is the topmost reason of most cancers death in human beings international. The dying tempo can be decreased if humans go for early analysis in order that suitable remedy can be administered by way of the clinicians inside exact time. In short, the cancer is an uncontrolled abnormal boom of unusual cells and invades the encircling tissues. Lung cancer can be in addition generalized in subsections, the first one is non-small cell lung cancer (NSCLC) and the second is small mobile lung cancer (SCLC). In this paper, the paintings is primarily based on NSCLC patients on account that it is more complicated and difficult to treatment. There are many contrasts regarding the detection and remedy of SCLC and NSCLC. There are diverse methods to locate the lung most cancers, one in all them is to apply its datasets except SVM and LR algorithms to constructed and increase the category and prediction version. Discovering the understanding from incredible datasets are regularly worn for records mining strategies. It has determined its crucial preserve in every pasture inclusive of fitness care. It has performed a chief position for extracting the hidden facts in the medical databases. The mining technique is more than the statistics evaluation which includes type, clustering, union law of mining and prediction. If the cancer has unfold, someone may also experience signs in different places within the anatomy. Its signs and symptoms are used to calculate the chance stage of ailment.

A lot of signs and symptoms are recognized on the premature segment. They are a pain within the chest, cough may be chronic, dry with phlegm, common breathing infections, shortness of breath, fatigue or loss of appetite, chest malaise and hoarseness. The primary consciousness of this examine is to predict the All over the world the sudden deaths are caused accidental and non-accidental ways. Non- accidental includes death by many causes such as disease. Cancer is one of the life threatening diseases, which is one of the major causes for non- accidental deaths. Lung cancer proved to be one of the major death for humans among all other cancers and it is also proved that early detection of lung cancer can be cured and increase human life. Cancer treatment depends on the stage that the patient is suffering, it may extend to chemotherapy, radiotherapy or surgery. Similarly, the patient's survival is based on the stage of cancer that surviving, the patient's with lung cancer is diagnosed 14% and they survive five years only.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405440

II. LITERATURE SURVEY

This chapter gives the overview of literature survey. This chapter represents some of the relevant work done by the researchers. Many existing techniques have been studied by the researchers on lung cancer problem, few of them are discussed below.

In [1], the author studied though imaging techniques. They trained the dataset from kaggle bowel 2017 data using convolutional neural network (CNN) algorithm. The author also used SVM classification for classification as Model A or Model B. Through the CT image scan, they segmented the portion of lung cancer and extracted the 15 features as text value and used for SVM algorithm. They found this method achieves good results on classification.The authors in paper [2] defined the work on early stage detection though unique strategy called using Ant lion optimizer plate detection from images. The author used optimization technique for the input feature before sending to the machine learning part. The author compared the work with Genetic algorithm, wolf optimization and Ant lion optimization, from these they concluded that ALO method outperforms in terms of accuracy of detection.

The author in [3] proposed lung disease prediction through rule based and classification techniques. For data preprocessing, the author applied One Dependency Augmented Naïve Bayes classifier (ODANB) and naive creedal classifier 2 (NCC2), the aim of the author is to exploit the use of hidden patterns for classification. The author experimented ODANB and proved that, it get better accuracy more than 70%, and the author also proved that it determines early stage of cancer through symptoms and patient details.

In [4], the author discussed logistic regression and SVM for lung cancer prediction. Through their experimental studied various evaluation metrics including true positive, true negative and false positives and false negatives. Through experimental results it is proved that logistic regression is achieving more accuracy than SVM algorithm. This work motivates us to do regression research than classification and clustering.

In [5], the author discussed about license plate detection for Chinese plates, they used Back propagation Neural network (BPNN) for character recognition. The author specifically designed the Chinese plate detection with specific length and width parameter. For identifying characters they used BPNN, which is trained with 50 epoch. However their model is good, but applicable only to Chinese number plates.

III. EXISTING SYSTEM

Existing imaging techniques are employed. The dataset was trained using the 2017 Kaggle bowel data with a convolutional neural network (CNN) algorithm. Additionally, the author implemented SVM classification to categorize the data as Model A or Model B. The Ant Lion Optimizer is utilized for detecting lung cancer from images. The author applied an optimization technique to the input features prior to processing them through the machine learning component. A comparison was made with Genetic Algorithms, Wolf Optimization, and Ant Lion Optimization, leading to the conclusion that the ALO method excels in detection accuracy. Classification methods employed include the One Dependency Augmented Naïve Bayes classifier (ODANB) and the Naive Creedal Classifier 2 (NCC2). The author's objective is to leverage hidden patterns for effective classification. Logistic regression and SVM are used for predicting lung cancer. Lung cancer prediction is conducted using the Artificial Neural Networks (ANN) algorithm, which exploits the use of macro neural networks, specifically a simple neural network without hidden layers to mitigate the over fitting issue..

IV.PROPOSED SYSTEM

The proposed system investigates the detection of lung cancer through the application of various regression algorithms. We have gathered the dataset for analysis from the UCI repository and utilized machine learning algorithms, including regression, for predicting lung diseases. There are three labels designated as 1, 2, and 3, which indicate the severity of the disease in an ascending manner. This scenario is regarded as a multi-class problem. Logistic regression, Linear Regression, Lasso, and Random Forest regression are employed to forecast lung diseases.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405440

V. METHODOLOGY

Regression techniques

Regression techniques are a class of supervised machine learning algorithms used to model and analyze the relationship between a dependent (target) variable and one or more independent (input) variables. These algorithms aim to predict continuous outcomes, such as tumor size, disease progression, or survival time in medical applications like lung cancer detection. Common regression methods include linear regression for simple linear relationships, polynomial regression for capturing non-linear trends, and regularized models like Lasso and Ridge regression to handle high-dimensional data and prevent overfitting. Advanced techniques such as Support Vector Regression (SVR), Random Forest Regression, and Gradient Boosting (e.g., XGBoost) offer robust solutions for complex and non-linear medical data. By accurately mapping input features to continuous outcomes, regression techniques play a crucial role in prognosis, treatment planning, and personalized healthcare.

MODULES

The modules included in our implementation are as follows

- ❖ Dataset collection
- ❖ Data pre-processing
- ❖ Training and prediction using Regression Models

DATASET COLLECTION

UCI repository data for lung cancer is taken for study. This contains around 32 samples with 56 attributes and with label. There are three labels named 1, 2 and 3, which represents the severity of disease in ascending order. This is considered to be a multi-class problem.

The dataset variable names are described below

Variable name	Attribute Description
Class	Class value labelled as 1,2 and 3
56 Attributes	Total 56 predictive attributes are considered

DATA PREPROCESSING

The dataset is split into train dataset and test dataset. For training purposes the X_Train attributes considered are 56 attributes from dataset and Y_train attribute is label. We considered the following phase of work in proposed model.

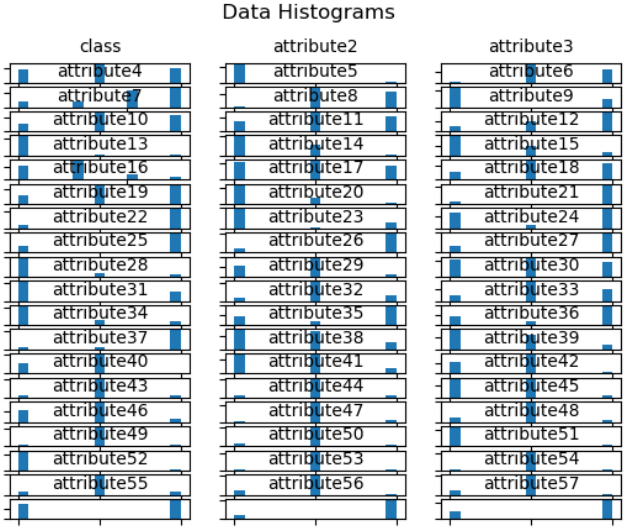


Figure: Visualization of Dataset as Histogram

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405440

TRAINING AND PREDICTION USING REGRESSION MODELS

Logistic regression

Logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

VI. EXPERIMENTAL RESULTS

The proposed work is implemented in Python 3.6.4 with libraries scikit-learn, pandas, matplotlib and other mandatory libraries. In our implementation Lung cancer prediction is done using machine learning algorithm Linear regression, Logistic regression, Lasso Regression and Random Forest Regression. The result shows that our proposed model achieves high accuracy using Random Forest Regression. The following table shows the implemented algorithm and its accuracy arrived.

Algorithm	Accuracy
Lasso Regression	50%
Logistic Regression	50%
Linear Regression	50%
Random Forest Regression	75%

The following chart represents the performance of proposed algorithm over considered dataset. Error Evaluations are carried out in the proposed system for MAE, MSE, RMSE and R-Squared are referred below.

Root-mean-square error (RMSE)

Root-mean-square error (RMSE) is a frequently used measure of the differences between values predicted by a model or an estimator and the values actually observed.

R-squared value

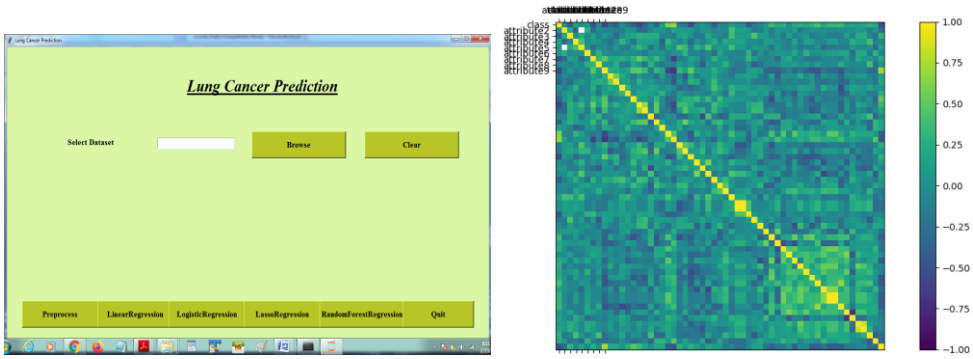
R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression.

Mean Absolute Error (MAE)

The Mean Absolute Error (or MAE) is the sum of the absolute differences between predictions and actual values. It gives an idea of how wrong the predictions were. The measure gives an idea of the magnitude of the error, but no idea of the direction (e.g. over or under predicting).

The Mean Squared Error (MSE)

The Mean Squared Error (or MSE) is much like the mean absolute error in that it provides a gross idea of the magnitude of error.



The following screen shows the Evaluation metrics for Linear regression algorithm for lung cancer prediction

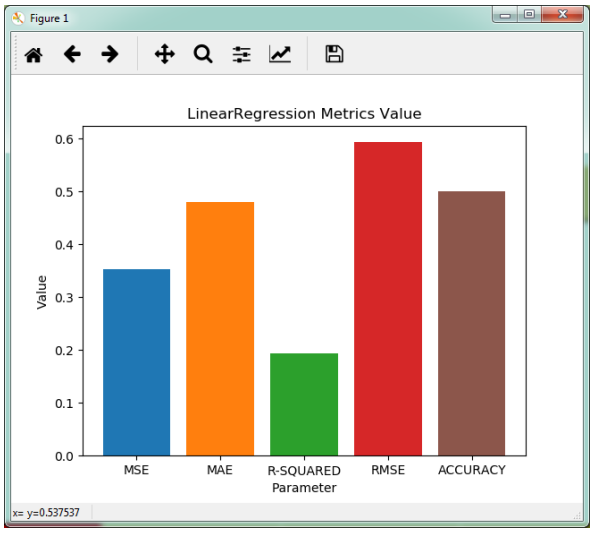
International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

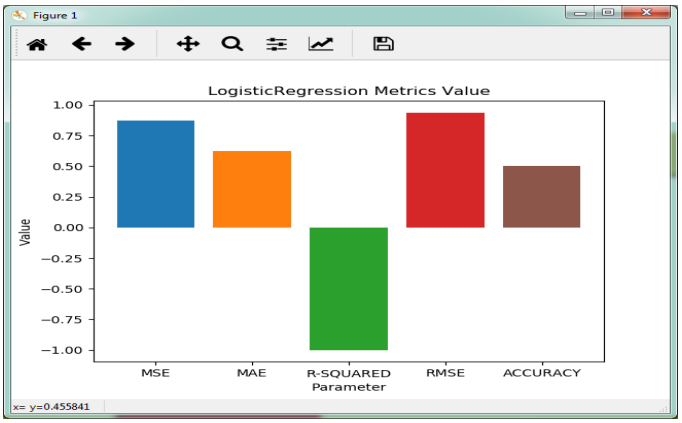
|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405440



The following screen shows the Evaluation metrics for Logistic regression algorithm for lung cancer prediction



The following screen shows the Evaluation metrics for Lasso regression algorithm for lung cancer prediction

The following screen shows the Evaluation metrics for Random Forest regression algorithm for lung cancer prediction

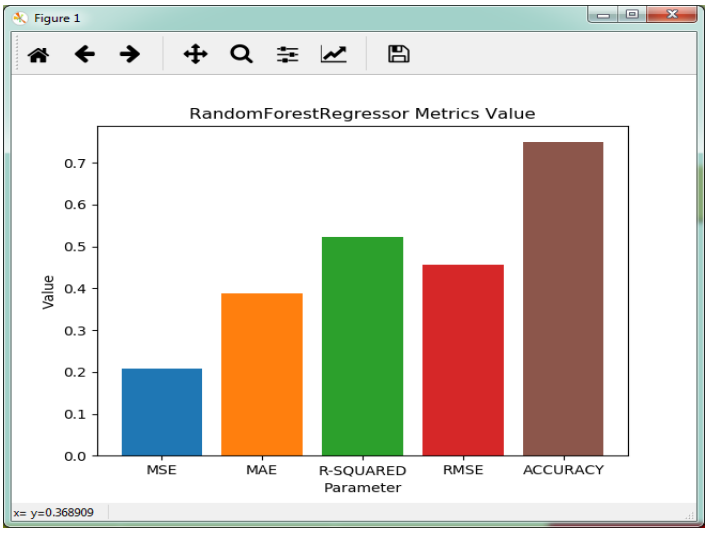
International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405440



VIII. CONCLUSION

Lung cancer is one of the major life-threatening diseases globally, primarily caused by lifestyle changes and the fact that individuals reside in more polluted areas. The incidence of deaths due to lung cancer is significantly high worldwide. Another contributing factor to these fatalities is the late-stage detection, which often results in loss of life. Early detection has the potential to save patients' lives. Therefore, there is a pressing need for a more effective system that can identify lung cancer at an early stage and protect lives from this illness. In the proposed system, a regression technique has been utilized to predict lung cancer. Our proposed system includes an experimental evaluation of metrics for Lasso regression, Logistic regression, Linear regression, and Random forest regression. The evaluation metrics determined include Accuracy, MAE, MSE, RMSE, and R-Squared error values. Our experiments demonstrate that the Random forest regression technique achieves a high accuracy rate of approximately 75% when compared to other regression techniques. The future scope of this work may involve the application of various deep learning models to attain higher detection rates.

REFERENCES

[1] F. Bray, et al., ”Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortalityworldwide for 36 cancers in 185 countries,” CA: A Cancer Journal for Clinicians, vol. 68, no. 6, pp. 394–424, 2018.
[2] M. Oudkerk, et al., ”European position statement on lung cancer screening,” The Lancet Oncology, vol. 18, no. 12, pp. e754–e766, 2017.
[3] A. Esteva, et al., ”A guide to deep learning in healthcare,” Nature Medicine, vol. 25, no. 1, pp. 24–29,2019.
[4] G. Litjens, et al., ”A survey on deep learning in medical image analysis,” Medical Image Analysis, vol. 42,pp. 60–88, 2017.
[5] N. Seah, ”Predicting Stroke Risk in Migraine Patients Using AI,” Arizona State Uni- versity, 2023.
[6] D. Wu, Data Mining with Python - Theory, Application, and Case Studies, CRC Press, 2022.
[7] G. Rana, R.K. Tailor, and V. Abdullayev, ”Data-Centric AI Solutions and Emerging Trends,” in Proc. IEEEConf. on AI in Medicine, 2022, pp. 421–428.
[8] C. Lo´pez-Camarillo, L.A. Marchat, et al., ”MicroRNAs in Cancer,” CRC Press, 2022.
[9] H.L. Gururaj, F. Flammini, V.R. Kumar, and N.S. Prema, ”Recent Trends in Health- care and AI,” in Proc.Int. Conf. on AI in Healthcare, 2021, pp. 233–241.
[10] C.M. Bishop, Pattern Recognition and Machine Learning. Springer, 2006.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details