

Performance Analysis of Machine Learning Models in Lung Cancer Detection

Jayalaxmi.Anem¹, D.V.L.N. Sastry², K. Manasa³, G. Hareesh⁴, V. Sai Saujanya⁵ and P. Prathyusha⁶

¹⁻⁶Aditya Institute of Technology and Management, Tekkali, AP, India-532201

Email: koradamanasa04@gmail.com, gujjidihareesh@gmail.com, veankumanthisaisowjanya@gmail.com, prathyushapanduru@gmail.com

Abstract— Lung cancer is one of the most life-threatening diseases, and early identification is key to improving survival rates. Advanced machine learning techniques offer a promising approach for predicting its occurrence. Various models, including Logistic Regression (LR), Random Forest (RF), Gaussian Naïve Bayes (GNB), K-Nearest Neighbors (KNN), Decision Tree (DT), and Support Vector Machine (SVM), were analyzed for their effectiveness in prediction. A dataset containing 309 instances with 16 attributes was utilized for evaluation. Metrics such as accuracy, precision, recall, and F1-score were used to measure performance. Among the models tested, logistic regression demonstrated the highest predictive capability.

Index Terms—Machine learning, Lung cancer detection, logistic regression, Predictive Model.

I. INTRODUCTION

Lung cancer remains a leading cause of death, with early detection being crucial for better survival rates. Machine learning (ML) offers a promising approach to improving diagnostic accuracy. Various models, including Logistic Regression (LR), Decision Tree (DT), k-Nearest Neighbors (KNN), Gaussian Naïve Bayes (GNB), Support Vector Machine (SVM), and Random Forest (RF), are evaluated using a dataset of 309 samples with 16 features. Performance metrics such as accuracy, precision, recall, and F1-score determine effectiveness. Identifying the most accurate model can enhance early detection and improve patient outcomes.

A. Classification Algorithms

Machine learning (ML) is the study and development of statistical algorithms that, without explicit programming, can recognize patterns in data and extrapolate those patterns to new data and tasks. Various machine learning techniques have been applied in every field in which we live, including word processing, computer vision, audio, email recognition, and even applications in agriculture and healthcare. Moreover, mathematical programming or optimization methods form the basis of many machine learning methodologies. Broadly speaking, the types of education fall into the following categories: The size of the problem and the data that machine learning algorithms use will determine which algorithms perform better in supervised learning than in unsupervised learning.

Logistic Regression

Logistic regression predicts binary outcomes by applying a threshold, assigning 1 for values above it and 0 for those below.

It works with discrete outputs and assumes multicollinearity among input variables. It's commonly used for tasks like text classification and credit risk evaluation.

K-Neighbors Classifier

The K-Nearest Neighbor (KNN) Classifier predicts a data point's category based on the majority vote of its closest neighbors. The number of neighbors, 'k', is usually set to $\sqrt{n}$ and adjusted if even to prevent ties.

Gaussian Naïve Bayes

Numerous uses for a probabilistic classifier exist, including vehicle control and medical diagnosis. GNB applies Bayes' rules under the assumption that the class's characteristics are independent model enables us to classify data effectively. Where "X" denotes a class, "Y" a specific object, and $P(X|Y)$ is calculated, it describes the density function that is employed in the model.

Support Vector Machine (SVM)

Support Vector Machine (SVM) classifies data by finding the best hyperplane that separates two groups, maximizing the margin with support vectors. It's used in tasks like handwriting recognition but can be slow with large datasets and sensitive to noise.

Random Forest

Random Forest reduces overfitting by using multiple models and majority voting for predictions. It's used in fraud detection and recommendations but can be memory-heavy, less interpretable, and less accurate than gradient boosting on complex tasks.

Decision Tree Algorithm

A Decision Tree splits data into subsets based on feature values, creating a tree structure. It's easy to understand and works for both regression and classification, but it may overfit without proper pruning or regularization.

II. MATERIALS AND METHODS

This section describes the data, methods, and approach employed in the analysis:

- Data and Libraries
- Data Cleaning and Preparation
- Feature Extraction and Selection
- Model Development

A. Data and libraries

The dataset contains 309 samples with 16 features, including clinical and demographic data like age, gender, smoking history, and genetic factors. It combines numerical and categorical variables, selected for their strong correlation with lung cancer, to ensure predictive accuracy.

TABLE I: INFORMATION ABOUT THE DATASET

S. No	Column	Non-null	Data type
0	Gender	309 non-null	int64
1	Age	309 non-null	int64
2	Smoking	309 non-null	int64
3	Yellow fingers	309 non-null	int64
4	Anxiety	309 non-null	int64
5	Peer pressure	309 non-null	int64
6	Chronic disease	309 non-null	int64
7	Fatigue	309 non-null	int64
8	Allergy	309 non-null	int64
9	Wheezing	309 non-null	int64
10	Alcohol consuming	309 non-null	int64
11	Coughing	309 non-null	int64
12	Shortness of breath	309 non-null	int64
13	Swallowing difficulty	309 non-null	int64
14	Chest pain	309 non-null	int64
15	Lung_cancer	309 non-null	int64

Columns were reviewed and processed to ensure no null values. The gender column was converted to binary (0 for female, 1 for male), and the "lung cancer" column was changed to numerical values (1 for positive, 0 for negative). Other categorical variables were standardized with binary values (1 for yes, 0 for no).

TABLE II: A TABLE SHOWING THE SYMPTOMS OF SIX PATIENTS (P1–P6) WITH VALUES 1 FOR NO AND 2 FOR YES.

	P1	P2	P3	P4	P5	P6
Gender	M	M	F	M	F	F
Age	69	74	59	63	63	72
Smoking	1	2	1	2	1	1
Yellow_fingers	2	1	1	2	2	2
Anxiety	2	1	1	2	1	1
Peer_pressure	1	1	2	1	1	1
Chronic disease	1	2	1	1	1	2
Fatigue	2	2	2	1	1	2
Allergy	1	2	1	1	1	2
Wheezing	2	1	2	1	2	2
Alcohol consuming	2	1	1	2	1	1
Coughing	2	1	2	1	2	2
Shortness of breath	2	2	2	1	2	2
Swallowing difficulty	2	2	1	2	1	1
Chest pain	2	2	2	2	1	1
Lung_cancer	YES	YES	NO	NO	NO	YES

B. Data Cleaning and Preparation

The data, with 309 rows and 16 columns, was preprocessed to remove bias and ensure model efficiency. Missing values were absent, gender and lung cancer columns were label-encoded, and irrelevant features were removed using correlation mapping.

C. Feature Extraction and Selection

Age, gender, smoking, and dyspnea were excluded from feature selection due to weak or negative correlations. Smoking's low predictive power may result from sample bias or variable interactions, with potential hidden patterns to be uncovered through further analysis.

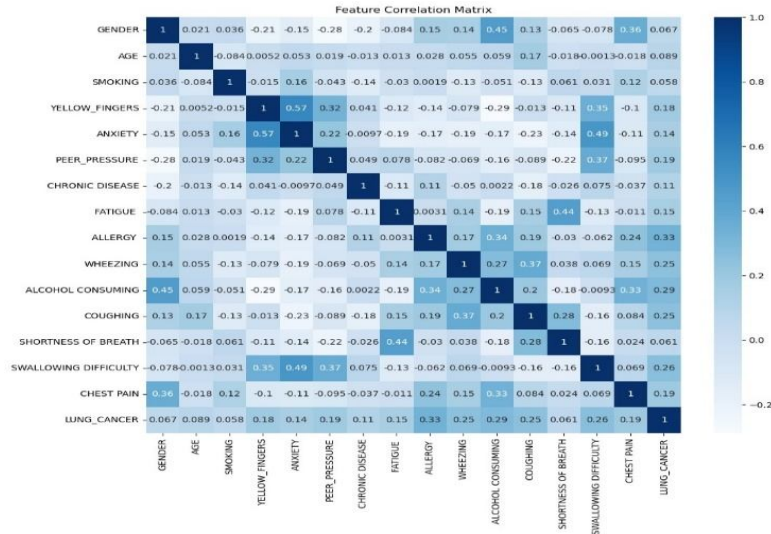


Figure 1: Feature correlation matrix

D. Model Development

Six machine learning algorithms were employed for prediction: Random Forest, Support Vector Machines, Gaussian Naive Bayes, K-Nearest Neighbors, Decision Trees, and Logistic Regression.

III. PROPOSED METHOD

The process begins with data analysis, including statistical methods, visualizations, and outlier detection. Feature selection eliminates irrelevant features to prevent overfitting and enhance performance.

The choice, training, and assessment of appropriate machine learning algorithms are guided by metrics such as accuracy, precision, recall, and F1 score. A comparative evaluation follows to determine the best model, which is then refined and deployed.

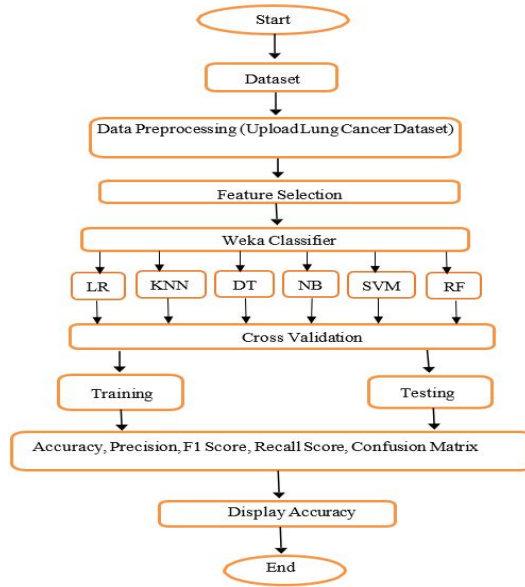


Figure 2. Overall study development process

VI. OBSERVATIONS

To identify where misclassifications occur, additional techniques are applied. Quality metrics such as precision, recall, F1-score, sensitivity, and specificity are calculated using the following formulas: False Positives (FP), False Negatives (FN), True Positives (TP), and True Negatives (TN).

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FN+FP} \quad [1]....$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad [2]....$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad [3]....$$

$$\text{F-score} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}} \quad [4]....$$

TABLE III: THE ACCURACY OF EVERY ALGORITHM

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.9909	0.9794	0.9896	0.9845
Decision Tree	0.9709	0.9794	0.9896	0.9845
(KNN)	0.9515	0.9691	0.9792	0.9741
Naïve Bayes	0.9515	0.9691	0.9792	0.9741
Random Forest	0.9709	0.9794	0.9896	0.9845
SVM	0.9515	0.9691	0.9792	0.9741

After training, the algorithms LR, DT, KNN, GNB, SVM, and RF achieved accuracy rates of 99%, 97%, 95%, 95%, 95%, and 97%, respectively (Table 3).

The confusion matrix is key for evaluating model performance, especially in imbalanced datasets, by revealing errors like false positives and false negatives. It helps fine-tune thresholds, optimize precision-recall, and improve models, making it essential for tasks like fraud detection and medical diagnosis.

V. RESULTS AND DISCUSSION

A. Logistic Regression

For binary classification and the prediction of lung cancer, logistic regression (LR) yields probabilistic results. It functions well with structured medical data and is easy to understand. It was the most successful model in the study, achieving the highest accuracy of 99%.

Confusion Matrix

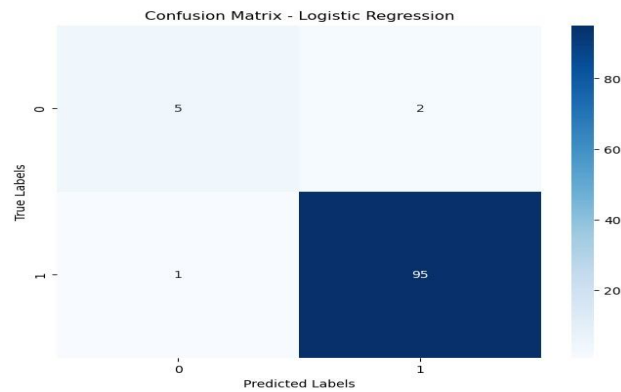


Fig 3: Confusion Matrix for Logistic Regression

B. Decision Tree

A Decision Tree (DT) simplifies the interpretation of data for the prediction of lung cancer by dividing it according to feature importance. It functions best when there are well-defined criteria for making decisions based on symptoms. However, it is prone to overfitting, particularly when dealing with noisy or small datasets.

Confusion Matrix

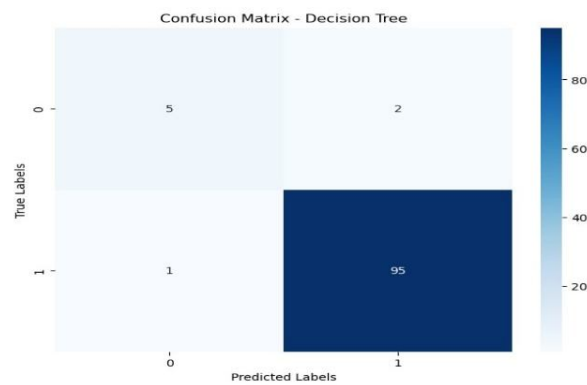


Fig 4: Confusion Matrix for Decision Tree

C. K-Nearest Neighbors (KNN)

It is easy to use but computationally costly for large datasets, and it may have trouble with unbalanced data. It classifies cases based on their similarity to neighboring data points.

Confusion Matrix

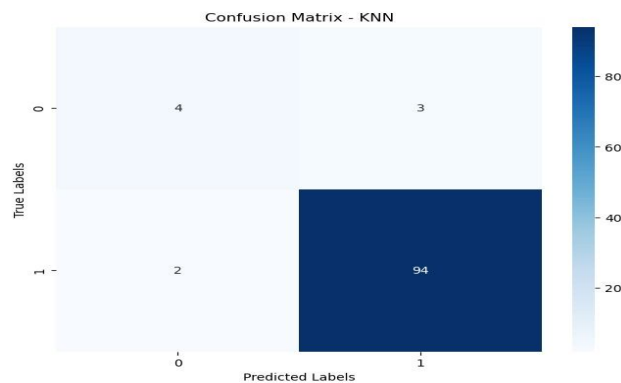


Fig 5: Confusion Matrix for KNN

D. Gaussian Naive Bayes (GNB)

It works well for probabilistic classifications and small datasets because it assumes feature independence; however, it suffers when features are highly correlated.

Confusion Matrix

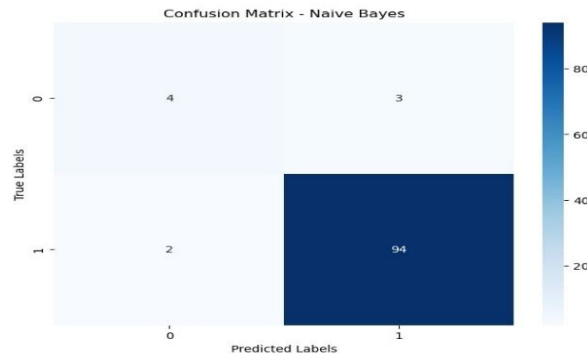


Fig 6: Confusion Matrix for Naive Bayes

E. Support Vector Machine (SVM)

Although it is computationally demanding and might not scale effectively for large datasets, it finds the optimal hyperplane to divide classes and performs well with high-dimensional data.

Confusion Matrix

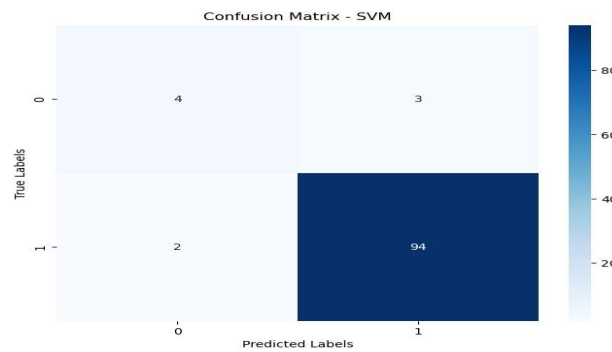


Fig 7: Confusion Matrix for SVM

F. Random Forest (RF)

Using multiple decision trees improves generalization and makes it resistant to overfitting, but it takes more processing power and storage space than simpler models.

Confusion Matrix

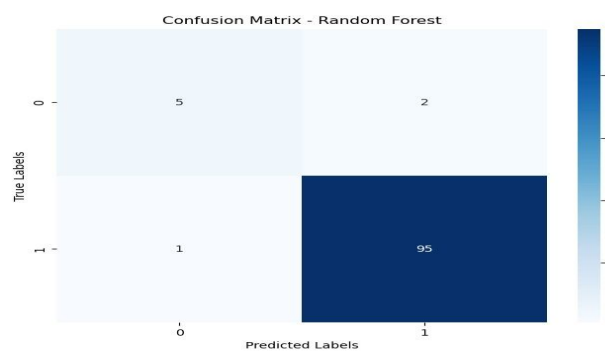


Fig 8: Confusion Matrix for Random Forest

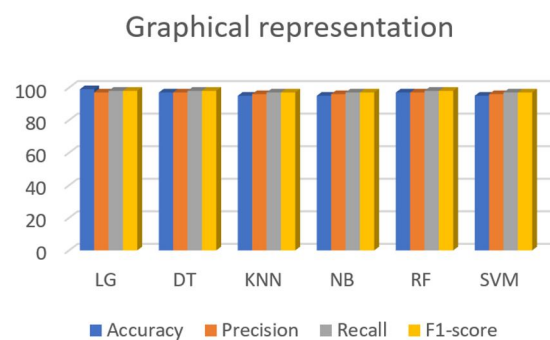


Fig 9: Graphical representation of all algorithms for predicting lung cancer

The graph illustrates a comparison of accuracy, precision, recall, and F1-score across various algorithms (LG, DT, KNN, NB, RF, SVM) in predicting lung cancer. With little preprocessing, logistic regression yields probabilistic results and is easy to use. It assumes a linear relationship, which reduces its effectiveness for complex data, but it trains rapidly without feature scaling. It has trouble with outliers, high-dimensional features, and big datasets without regularization.

TABLE IV: FINDINGS FOLLOWING TESTING OF THE MODEL CONSTRUCTED WITH THE LOGISTIC REGRESSION ALGORITHM

Case	P1	P2	P3	P4	P5
Yellow_fingers	No	No	No	No	Yes
Anxiety	Yes	No	No	Yes	Yes
Peer_pressure	No	Yes	No	No	Yes
Chronic disease	No	Yes	Yes	No	No
Fatigue	Yes	Yes	Yes	No	No
Allergy	No	No	No	Yes	Yes
Wheezing	Yes	No	No	Yes	No
Alcohol consuming	No	No	No	Yes	Yes
Coughing	Yes	Yes	Yes	Yes	No
Swallowing difficulty	No	No	No	No	No
Chest pain	Yes	Yes	No	No	No
Lung_cancer	NO	YES	NO	YES	YES

VI. CONCLUSION

Logistic Regression, K-Nearest Neighbors, Gaussian Naïve Bayes, Random Forest, Support Vector Machine, and Decision Tree were evaluated for their ability to predict lung cancer. After thorough testing, Logistic Regression proved to be the top-performing model, achieving the highest accuracy of 99%. Although KNN, SVM, and Naïve Bayes had slightly lower accuracy, models like Decision Tree and Random Forest also delivered strong results.

REFERENCES

- [1] Negar Maleki, "A k-NN method for lung cancer prognosis with the use of a genetic algorithm for feature selection," in 2020 Expert Systems With Applications, ESWA 2020, P.O. Box 11155-9414, doi:https://doi.org/10.1016/j.eswa.2020.113981
- [2] Kapila Moon1, Ashok Jetawat, "Predicting Lung Cancer with K-Nearest Neighbors (KNN): A Computational Approach," in Proceedings - 2024 Indian Journal Of Science And Technology, Udaipur, Rajasthan, India, 2024. doi:https://doi.org/10.17485/IJST/v17i21.1192
- [3] B. Ramu, Srija Vakiti, Vaishnavi Asuri, "Analysis of lung cancer prediction using machine learning algorithms," -2022 International Journal of Health Sciences, Changsha, 8623–8634. doi: https://doi.org/10.53730/ijhs.v6nS3.8053
- [4] Joselyn Zapata-Paulini, Michael Cabanillas-Carbonell "Performance analysis of 10 machine learning models in lung cancer prediction," -International Conference on Indonesian Journal of Electrical Engineering and Computer Science, Lima, Peru, 2024, pp1352-1364, doi: https://doi.org/10.11591/ijeecs.v37.i2.pp1352-1364
- [5] Yawei Li, Xin Wu, Ping Yang, Guoqian Jiang, Yuan Luo, "Machine Learning for Lung Cancer Diagnosis, Treatment, and Prognosis," -2022 Genomics Proteomics Bioinformatics, Chikkamagaluru, Chicago, USA, 2022, doi: https://doi.org/10.1016/j.gpb.2022.11.003
- [6] Sarreha Tasmin Rikta, Khandaker Mohammad Mohi Uddin, Nitish Biswas, Rafid Mostafiz, Fateha Sharmin, Samrat Kumar Deyet. "XML-GBM lung: An explainable machine learning-based application for the diagnosis of lung cancer, 2023., doi: http://dx.doi.org/10.1016/j.jpi.2023.100307
- [7] Yang Yang, Li Xu, Liangdong Sun, Peng Zhang, Suzanne S. Farid "Machine learning application in personalised lung cancer recurrence and survivability prediction", 2022, doi: https://doi.org/10.1016/j.csbj.2022.03.035
- [8] Kamta Nath Mishra, Alok Mishra, Soumya Ray, Anjali Kumari, Saad Misbah Waris "Enhancing cancer detection and prevention mechanisms using advanced machine learning approaches-", 2024. doi: https://doi.org/10.1016/j.imu.2024.101579
- [9] Fatimah Abdulazim Altuhaifa, Khin Than Win, Guoxin Su "Predicting lung cancer survival based on clinical data using machine learning: A review", 2023, doi: https://doi.org/10.1016/j.compbiomed.2023.107338
- [10] Mahmood Khalsan "A Survey of Machine Learning Approaches Applied to Gene Expression Analysis for Cancer Prediction", 2022, IEEE Access, U.K., doi: https://doi.org/ 0.1109/access.2022.3146312