International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 08, August 2026)

Lung Cancer Prediction Using Machine Learning Techniques for Early Diagnosis

Dr. S. Pavani

Assistant Professor, Department of Computer Science, C. M. Dubey P.G. College, Bilaspur, Chhattisgarh, India

***Abstract*-- Lung cancer is one of the most common and deadliest cancers worldwide, accounting for a significant number of cancer-related deaths each year. Early detection plays a crucial role in improving survival rates and reducing mortality. Recent advancements in Machine Learning (ML) have enabled the development of intelligent diagnostic systems capable of predicting lung cancer using patient symptoms and clinical data. This study investigates the performance of various machine learning algorithms, including Logistic Regression, Support Vector Machine (SVM), Decision Tree, Random Forest, Gradient Boosting, and XGBoost for lung cancer prediction. The dataset used contains 309 patient records with multiple risk factors and symptoms associated with lung cancer. Experimental results indicate that Random Forest, Gradient Boosting, and XGBoost classifiers achieved the highest prediction accuracy of 92.31%. The findings demonstrate the effectiveness of machine learning techniques in supporting healthcare professionals for early lung cancer diagnosis and treatment planning.**

***Keywords*--Lung Cancer Prediction, Machine Learning, Random Forest, XGBoost, Gradient Boosting, Healthcare Analytics, Artificial Intelligence.**

I. Introduction

Lung cancer remains one of the leading causes of cancer-related mortality worldwide. According to global health statistics, millions of new cases are diagnosed annually, with a high mortality rate due to late-stage detection. Smoking, environmental pollution, genetic predisposition, and occupational hazards are among the major risk factors contributing to lung cancer.

Early diagnosis significantly improves treatment outcomes and survival rates. Traditional diagnostic procedures such as CT scans, biopsies, and laboratory examinations are often expensive and time-consuming. Artificial Intelligence (AI) and Machine Learning (ML) provide promising solutions for developing predictive models that can assist clinicians in identifying high-risk patients at an early stage.

Machine learning algorithms can analyze large datasets, recognize hidden patterns, and provide accurate predictions. This research focuses on comparing different machine learning techniques for lung cancer prediction using patient symptom data.

II. Literature Review

Several researchers have explored machine learning techniques for cancer diagnosis and prediction.

Roy et al. utilized image processing and machine learning techniques for lung cancer detection using CT scan images. Their study employed Support Vector Machine (SVM) and Random Forest algorithms, achieving an accuracy of approximately 94.5%.

Faisal et al. compared Multilayer Perceptron (MLP), Naïve Bayes, Decision Tree, Neural Networks, Gradient Boosting, and SVM classifiers for lung cancer prediction. Their findings revealed that Gradient Boosting outperformed other classifiers with an accuracy of nearly 90%.

Recent studies have demonstrated that ensemble learning algorithms such as Random Forest and XGBoost provide superior performance due to their ability to handle complex nonlinear relationships and reduce overfitting.

Despite significant progress, comparative evaluations of multiple machine learning algorithms using symptom-based datasets remain limited. Therefore, this study aims to analyze and compare various classification techniques for effective lung cancer prediction.

III. Dataset Description

The dataset used in this study was obtained from Kaggle and contains clinical information associated with lung cancer prediction.

Dataset Characteristics

- Total Records: 309
- Total Attributes: 16
- Target Variable: Lung Cancer (Yes/No)

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 08, August 2026)

Important Features

Feature	Description
Age	Age of patient
Gender	Male/Female
Smoking	Smoking habit
Yellow Fingers	Tobacco staining indicator
Anxiety	Anxiety level
Peer Pressure	Social influence
Chronic Disease	Existing chronic illness
Fatigue	Fatigue symptoms
Wheezing	Breathing difficulty
Alcohol Consumption	Alcohol intake
Coughing	Persistent cough
Shortness of Breath	Respiratory difficulty
Swallowing Difficulty	Dysphagia
Chest Pain	Pain in chest

The dataset provides a balanced representation of symptoms commonly associated with lung cancer diagnosis.

IV. Methodology

4.1 Data Preprocessing

Data preprocessing is essential to improve model performance. The following steps were performed:

- Data cleaning
- Missing value analysis
- Label encoding
- Feature selection
- Data normalization
- Train-test splitting

No missing values were found in the dataset.

4.2 Feature Selection

Eight significant symptoms were selected as predictive features:

1. Yellow Fingers
2. Anxiety
3. Fatigue
4. Wheezing
5. Coughing
6. Shortness of Breath
7. Swallowing Difficulty
8. Chest Pain

These symptoms were used as independent variables for model training.

4.3 Machine Learning Algorithms

The following classification models were implemented:

Logistic Regression

A statistical classification technique suitable for binary outcomes.

Support Vector Machine (SVM)

Finds the optimal hyperplane that maximizes class separation.

Decision Tree

Creates hierarchical decision rules for classification.

Random Forest

An ensemble learning method using multiple decision trees.

Gradient Boosting

Builds sequential weak learners to minimize classification errors.

XGBoost

An optimized gradient boosting framework offering high efficiency and accuracy.

V. Performance Evaluation Metrics

Model performance was evaluated using the following metrics:

Accuracy

[
Accuracy = \frac{TP + TN}{TP + TN + FP + FN}
]

Precision

[
Precision = \frac{TP}{TP + FP}
]

Recall (Sensitivity)

[
Recall = \frac{TP}{TP + FN}
]

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 08, August 2026)

F1-Score

[
F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}
]

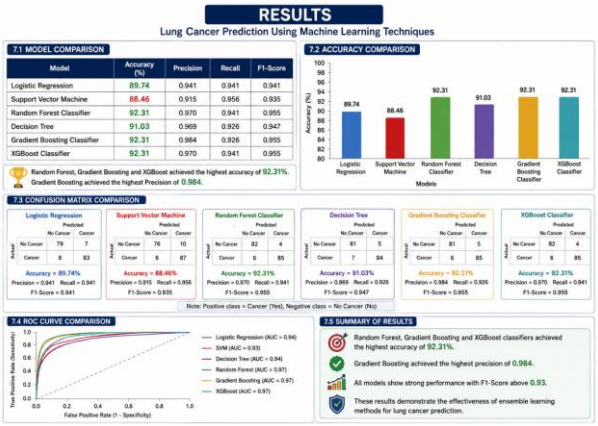
ROC-AUC

Measures the classifier's ability to distinguish between positive and negative classes.

VI. Results And Discussion

Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	89.74%	0.941	0.941	0.941
Support Vector Machine	88.46%	0.915	0.956	0.935
Decision Tree	91.03%	0.969	0.926	0.947
Random Forest	92.31%	0.970	0.941	0.955
Gradient Boosting	92.31%	0.984	0.926	0.955
XGBoost	92.31%	0.970	0.941	0.955



Discussion

The results indicate that ensemble learning methods significantly outperform traditional classifiers. Random Forest, Gradient Boosting, and XGBoost achieved the highest accuracy of 92.31%.

Gradient Boosting demonstrated the highest precision value (0.984), indicating a lower false-positive rate. XGBoost and Random Forest also showed strong predictive capabilities due to their ensemble architecture.

The findings confirm that machine learning techniques can effectively assist healthcare professionals in identifying lung cancer at an early stage and improving diagnostic accuracy.

VII. Conclusion And Future Scope

This study presented a comparative analysis of six machine learning algorithms for lung cancer prediction using symptom-based clinical data. The performance evaluation demonstrated that Random Forest, Gradient Boosting, and XGBoost classifiers achieved the best results with an accuracy of 92.31%.

Machine learning-based diagnostic systems can support physicians in early disease detection, reduce healthcare costs, and improve patient survival rates.

Future Scope

Future research can focus on:

1. Integration of medical imaging data such as CT scans and X-rays.
2. Application of Deep Learning models such as CNN and LSTM.
3. Development of real-time clinical decision support systems.
4. Hyperparameter optimization for improved prediction accuracy.
5. Explainable AI techniques to improve model transparency in healthcare applications.

The integration of Artificial Intelligence with healthcare systems has the potential to revolutionize early cancer diagnosis and improve patient outcomes.

REFERENCES

[1] Bhat, S. M. Lung Cancer Dataset, Kaggle Repository.
[2] Brindha, A. A., Indirani, S., & Srinivasan, A. Lung Cancer Detection Using SVM Algorithm and Optimization Techniques, 2016.
[3] Kurkure, M., & Thakare, A. Automated System for Lung Cancer Detection Using Evolutionary Approach, 2016.
[4] Rani, B., Goel, A. K., & Kaur, R. Modified Approach for Lung Cancer Detection Using Optimization Algorithms, 2016.
[5] Fontana, R. S., et al. Screening for Lung Cancer. Cancer Journal, 1991.
[6] Henschke, C. I., et al. Early Lung Cancer Action Project. Lancet, 1999.
[7] Roy, S., et al. Lung Cancer Detection Using Image Processing and Machine Learning Techniques.
[8] Faisal, M., et al. Comparative Study of Machine Learning Algorithms for Lung Cancer Diagnosis.
[9] Deo, R. Machine Learning in Medicine. Circulation, 2015.
[10] Johnson, K. W., et al. Artificial Intelligence in Healthcare Diagnostics. Journal of Medical Systems, 2018.