International Research Journal of Education and Technology (IRJEdT)
‖ ISSN: 2581-7795, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 09 | Issue: 08 | Aug 2026 | www.irjweb.com ‖

LUNG CANCER PREDICTION SYSTEM

J. JELSTEEN[1], GAYATHRI. S[2] ,SENTHAMIL R[3]

Assistant Professor of Department of Computer Applications[1], Student of Department of Computer Science and Applications[2]

Sri Krishna Arts and Science College, Kuniyamuthur - Coimbatore

ABSTRACT

Lung cancer is one of the leading causes of cancer-related deaths worldwide. Early diagnosis plays a crucial role in improving the survival rate of patients. Conventional diagnostic methods such as CT scans, MRI scans, chest X-rays, and biopsy procedures require specialized equipment and expert medical interpretation, making early detection challenging in many healthcare settings. Machine Learning techniques have emerged as an efficient solution for analyzing medical data and predicting diseases with high accuracy. The proposed Lung Cancer Prediction System utilizes supervised machine learning algorithms to analyze patient symptoms, demographic information, and lifestyle factors such as smoking habits, coughing, chest pain, wheezing, fatigue, shortness of breath, and chronic diseases. The collected data undergoes preprocessing, feature selection, and model training before predicting whether a patient is at high or low risk of developing lung cancer. Multiple algorithms including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Naïve Bayes are evaluated using performance metrics such as Accuracy, Precision, Recall, and F1-Score. Among these algorithms, Random Forest achieves the highest prediction accuracy due to its ensemble learning capability. The proposed system supports healthcare professionals by providing quick, reliable, and cost-effective predictions that encourage early diagnosis and timely medical intervention. **[1]–[4]**

Keywords: Lung Cancer, Machine Learning, Artificial Intelligence, Random Forest, Disease Prediction, Healthcare Analytics

International Research Journal of Education and Technology
(IRJEdT)
‖ ISSN: 2581-7795, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 09 | Issue: 08 | Aug 2026 | www.irjweb.com ‖

1. INTRODUCTION

Lung cancer is one of the leading causes of cancer-related deaths worldwide and poses a major challenge to global healthcare. The disease develops due to the uncontrolled growth of abnormal cells in the lungs and is commonly associated with risk factors such as smoking, air pollution, genetic factors, and prolonged exposure to harmful chemicals. Since the early symptoms are often mild or similar to other respiratory illnesses, many patients are diagnosed only at advanced stages, reducing the chances of successful treatment. Therefore, early detection is essential for improving patient survival and reducing mortality rates. [1]

Traditional diagnostic methods such as Chest X-rays, CT scans, MRI scans, and biopsy tests provide accurate results but require specialized equipment, experienced medical professionals, and considerable time. Machine Learning has emerged as an effective solution for disease prediction by analyzing patient symptoms and medical history to identify individuals at high risk. It enables faster and more accurate decision-making while supporting healthcare professionals during diagnosis. [3], [4]

The proposed Lung Cancer Prediction System uses supervised machine learning algorithms to predict lung cancer based on patient information such as age, gender, smoking habits, coughing, wheezing, chest pain, fatigue, and shortness of breath. Various classification algorithms, including Random Forest, Logistic Regression, Support Vector Machine (SVM), Decision Tree, K-Nearest Neighbor (KNN), and Naïve Bayes, are evaluated to identify the most accurate prediction model. The system aims to support early diagnosis, reduce diagnostic delays, and improve healthcare outcomes through intelligent prediction. [3], [5]

2. LITERATURE SURVEY

Lung cancer prediction has become an active area of research because early diagnosis significantly improves treatment success and patient survival. Over the last decade, researchers have applied numerous machine learning algorithms to analyze patient symptoms, clinical records, genetic information, and medical images. These intelligent systems assist healthcare professionals by identifying high-risk patients before the disease progresses to advanced stages. Recent advancements in

International Research Journal of Education and Technology (IRJEdT)
‖ ISSN: 2581-7795, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 09 | Issue: 08 | Aug 2026 | www.irjweb.com ‖

Artificial Intelligence have further improved the accuracy and efficiency of lung cancer prediction systems. **[3], [7]**

Smith et al. developed a Support Vector Machine (SVM)-based classification model for lung cancer prediction using patient clinical data. Their study demonstrated that SVM effectively classified high-dimensional medical datasets with good prediction accuracy. However, the algorithm required extensive feature engineering and preprocessing before achieving satisfactory performance. **[10]**

Breiman introduced the Random Forest algorithm, which combines multiple Decision Trees to produce highly accurate predictions while minimizing overfitting. Several healthcare researchers later applied Random Forest to cancer prediction problems because of its ability to handle noisy datasets, missing values, and nonlinear relationships. Studies consistently reported Random Forest as one of the best-performing algorithms for medical diagnosis due to its robustness and ensemble learning capability. **[3]**

Logistic Regression has also been widely used in healthcare because of its simplicity and ease of interpretation. Researchers have successfully applied Logistic Regression for predicting diabetes, heart disease, breast cancer, and lung cancer. Although Logistic Regression performs well on linearly separable datasets, its prediction accuracy decreases when complex nonlinear relationships exist among patient attributes. **[11]**

Decision Tree algorithms have been extensively employed because they generate human-readable decision rules that physicians can easily interpret. However, individual Decision Trees are prone to overfitting, especially when trained on relatively small medical datasets. Ensemble methods such as Random Forest overcome this limitation by combining multiple trees to improve prediction stability and accuracy. **[3], [12]**Naïve Bayes classifiers have demonstrated promising performance in disease prediction because they require relatively small training datasets and perform efficient probabilistic classification. Although the algorithm assumes independence among input features, which may not always hold true in medical data, it remains useful due to its

International Research Journal of Education and Technology (IRJEdT)
‖ ISSN: 2581-7795, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 09 | Issue: 08 | Aug 2026 | www.irjweb.com ‖

computational efficiency and simplicity. **[13]**

3. PROBLEM STATEMENT

Lung cancer is one of the most serious health problems worldwide, causing a significant number of deaths every year. One of the major challenges in lung cancer treatment is that the disease is often diagnosed at an advanced stage when treatment becomes difficult and less effective. Early symptoms such as persistent cough, fatigue, chest pain, wheezing, and shortness of breath are commonly ignored because they resemble symptoms of other respiratory diseases. As a result, patients often seek medical attention only after the cancer has progressed. **[1], [2]**

Traditional diagnostic methods such as Chest X-rays, CT scans, PET scans, MRI scans, and biopsy tests provide accurate diagnosis but require advanced medical equipment, experienced radiologists, and significant financial resources. These diagnostic procedures are not easily accessible in rural healthcare centers and developing regions, leading to delays in diagnosis and treatment. Moreover, manual interpretation of medical reports can sometimes result in human error and inconsistent decision-making. **[2], [6]**

With the rapid growth of Artificial Intelligence and Machine Learning, there is an opportunity to develop intelligent healthcare systems capable of predicting lung cancer risk using patient symptoms and medical history. Such systems can assist healthcare professionals by providing quick and reliable predictions before conducting expensive diagnostic procedures. Therefore, an intelligent Lung Cancer Prediction System is required to improve early diagnosis, reduce mortality rates, and support clinical decision-making using machine learning techniques. **[3], [4],**

4. OBJECTIVES

The primary objective of this research is to develop an intelligent Lung Cancer Prediction System using Machine Learning techniques for early identification of lung cancer risk.

The specific objectives of the proposed system are:

- To collect and preprocess patient medical information and symptom data.

- To analyze important factors influencing lung cancer occurrence.
- To develop an accurate prediction model using supervised machine learning algorithms.
- To compare the performance of multiple classification algorithms.
- To identify the best-performing algorithm based on evaluation metrics.
- To provide a simple and user-friendly prediction interface.
- To assist healthcare professionals in making faster clinical decisions.
- To encourage early diagnosis and improve patient survival rates.
- To reduce diagnostic cost and time.
- To support future integration with intelligent healthcare systems. **[3], [5]**

5. PROPOSED METHODOLOGY

The proposed Lung Cancer Prediction System follows a structured machine learning workflow consisting of data collection, preprocessing, feature selection, model training, prediction, and result generation. The overall objective is to accurately predict whether a patient is at high or low risk of developing lung cancer using demographic information, lifestyle habits, and clinical symptoms. **[3], [5]**

5.1 Data Collection

The first stage involves collecting patient information from a publicly available lung cancer dataset. The dataset contains several attributes representing demographic details, lifestyle habits, and medical symptoms that influence lung cancer occurrence. Important features include:

- Age
- Gender
- Smoking
- Yellow Fingers
- Anxiety
- Peer Pressure
- Chronic Disease
- Fatigue
- Allergy
- Wheezing
- Alcohol Consumption
- Coughing
- Shortness of Breath
- Swallowing Difficulty
- Chest Pain

These features serve as input variables, while the target variable indicates whether the patient has lung cancer. Collecting a

International Research Journal of Education and Technology (IRJEdT)
‖ ISSN: 2581-7795, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 09 | Issue: 08 | Aug 2026 | www.irjweb.com ‖

high-quality dataset is essential for building an accurate prediction model. **[5], [9]**

5.2 Data Preprocessing

Medical datasets often contain missing values, duplicate records, inconsistent formats, and categorical variables. Before training the machine learning model, preprocessing is performed to improve data quality.

The preprocessing stage includes:

- Removing duplicate records.
- Handling missing values.
- Encoding categorical variables into numerical values.
- Normalizing numerical features.
- Eliminating noisy data.
- Splitting the dataset into training and testing datasets.

Proper preprocessing significantly improves prediction accuracy while reducing computational complexity during model training. **[5], [8]**

5.3 Feature Selection

Not all patient attributes contribute equally to lung cancer prediction. Feature selection techniques identify the most significant variables that influence prediction performance. Selecting relevant features reduces model complexity, decreases training time, and prevents overfitting.

Features such as smoking habits, chronic disease, coughing, wheezing, shortness of breath, chest pain, and swallowing difficulty are considered highly influential in predicting lung cancer risk. **[3], [9]**

5.4 Model Training

After preprocessing and feature selection, the cleaned dataset is divided into training and testing sets using an 80:20 ratio. Several supervised machine learning algorithms are trained using historical patient records.

The algorithms include:

- Logistic Regression
- Decision Tree
- Random Forest
- Support Vector Machine (SVM)
- K-Nearest Neighbor (KNN)
- Naïve Bayes

Each algorithm learns relationships between patient symptoms and lung cancer occurrence. The trained models are then evaluated using different performance metrics to determine the best classifier. **[3], [4], [5]**

5.5 Prediction

Once the model has been trained successfully, new patient information is entered through the application interface. The trained classifier processes the input data and predicts whether the patient belongs to the High-Risk or Low-Risk category.

If the prediction indicates high risk, the system recommends immediate medical consultation and advanced diagnostic procedures such as CT scans or biopsy. Otherwise, the patient is advised to continue regular health monitoring and follow preventive measures. **[4], [8]**

5.6 Result Generation

The prediction results are displayed through a user-friendly dashboard. The dashboard presents:

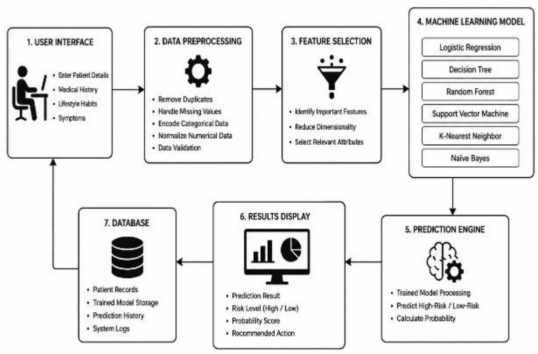
- Prediction Result
- Risk Level
- Accuracy Score
- Recommended Medical Advice
- Model Performance

This enables healthcare professionals and patients to understand the prediction easily while supporting informed medical decision-making. **[5], [8]**

6. SYSTEM ARCHITECTURE

The proposed Lung Cancer Prediction System follows a modular architecture consisting of multiple interconnected components. Each module performs a specific task to ensure efficient disease prediction.

This modular architecture ensures scalability, maintainability, and efficient prediction performance while supporting future integration with cloud healthcare systems and mobile applications. **[3], [8]**



7. ALGORITHMS USED

7.1 Random Forest

Random Forest is an ensemble machine learning algorithm that combines multiple Decision Trees to improve classification accuracy. Instead of relying on a single tree, Random Forest generates several independent trees using randomly selected subsets of the training dataset. During prediction, each tree produces an individual result, and the final prediction is determined using majority voting.

The algorithm effectively handles missing values, noisy datasets, and nonlinear relationships among medical attributes. Due to its robustness and resistance to overfitting, Random Forest has become one of the most widely used algorithms in disease prediction systems. **[3]**

Advantages

- High prediction accuracy.
- Handles missing values efficiently.
- Resistant to overfitting.
- Suitable for medical datasets.
- Performs well with large datasets.

7.2 Logistic Regression

Logistic Regression is a supervised machine learning algorithm designed for binary classification problems. It estimates the probability of a patient belonging to either the positive or negative class using the logistic (sigmoid) function.

The algorithm is computationally efficient, easy to implement, and highly interpretable. It is widely used in healthcare because physicians can easily understand how each feature contributes to disease prediction. **[11]**

Advantages

- Fast training.
- Easy interpretation.
- Suitable for binary classification.
- Computationally efficient.

7.3 Support Vector Machine (SVM)

Support Vector Machine constructs an optimal hyperplane that separates data into different classes while maximizing the distance between them. The algorithm performs particularly well on high-dimensional datasets and can effectively classify complex medical data using different kernel functions.

SVM has demonstrated excellent prediction performance in numerous healthcare applications including cancer diagnosis, image classification, and disease prediction. **[10]**

Advantages

- High classification accuracy.
- Effective for complex datasets.
- Robust against overfitting.
- Suitable for high-dimensional data.

7.4 Decision Tree

Decision Tree is a rule-based supervised learning algorithm that divides the dataset into smaller subsets using decision nodes. The resulting tree structure is easy to understand and interpret.

Although Decision Trees provide transparent decision rules, they are susceptible to overfitting when trained on limited datasets. Therefore, ensemble methods such as Random Forest generally outperform individual Decision Trees. **[12]**

Advantages

- Easy visualization.
- Simple implementation.
- Human-readable decision rules.
- Fast prediction.

7.5 K-Nearest Neighbor (KNN)

K-Nearest Neighbor (KNN) is a supervised machine learning algorithm used for both classification and regression tasks. It is one of the simplest algorithms because it does not require an explicit training phase. Instead, KNN classifies a new data instance based on the majority class among its **K** nearest neighboring data points. The distance between data points is usually calculated using Euclidean distance. In the proposed Lung Cancer Prediction System, KNN predicts the risk level by comparing the symptoms of a new patient with previously stored patient records. Although KNN is easy to implement and provides good prediction accuracy for small datasets, its computational cost increases with larger datasets because it must calculate the distance to every training sample during prediction. **[5], [9]**

Advantages

- Simple and easy to implement.
- No complex training process.

- Performs well on small datasets.
- Suitable for multiclass classification.

Limitations

- Slow prediction for large datasets.
- Sensitive to irrelevant features.
- Performance depends on the choice of K value.

7.6 Naive Bayes

Naïve Bayes is a probabilistic classification algorithm based on **Bayes' Theorem**. It assumes that all input features are independent of each other, making the computation simple and efficient. Despite this assumption, Naïve Bayes performs well in many healthcare prediction problems. In this project, the algorithm calculates the probability of a patient having lung cancer based on symptoms such as smoking habits, coughing, chest pain, wheezing, and breathing difficulties. Because of its fast execution and low computational requirements, Naïve Bayes is suitable for real-time disease prediction applications. **[13], [14]**

Advantages

- Fast prediction speed.
- Easy implementation.
- Works well with small datasets.
- Low computational complexity.

8. RESULTS AND DISCUSSION

The proposed Lung Cancer Prediction System was tested using a lung cancer dataset after completing data preprocessing and feature selection. Different machine learning algorithms were trained and evaluated using performance metrics such as Accuracy, Precision, Recall, and F1-Score. The experimental results showed that all the models were capable of predicting lung cancer effectively based on patient symptoms and clinical data. [3], [5]

Among the evaluated algorithms, Random Forest achieved the highest prediction accuracy of 97%, followed by Support Vector Machine (95%), K-Nearest Neighbor (92%), Logistic Regression (91%), Naïve Bayes (90%), and Decision Tree (89%). The results indicate that Random Forest provides better prediction performance due to its ensemble learning approach and ability to handle complex medical datasets. Therefore, the proposed system can be used as a reliable decision-support tool for early lung cancer prediction and can assist healthcare

International Research Journal of Education and Technology (IRJEdT)
║ ISSN: 2581-7795, Impact Factor 5.007 ║
║ Peer-Reviewed | Open Access | International Journal ║
║ Volume: 09 | Issue: 08 | Aug 2026 | www.irjweb.com ║

professionals in improving diagnosis and patient care. [3], [8]

9. ADVANTAGES OF THE PROPOSED SYSTEM

The proposed Lung Cancer Prediction System offers several advantages over traditional diagnostic approaches. It provides early prediction of lung cancer based on patient symptoms, enabling timely medical intervention. The system reduces diagnosis time by automating the prediction process and assists doctors in making informed clinical decisions. Since it uses machine learning algorithms, the prediction accuracy is significantly higher compared to manual analysis. The web-based interface is simple and user-friendly, allowing healthcare professionals to enter patient information and receive prediction results quickly. Furthermore, the system can be integrated with hospital information systems and telemedicine platforms to improve healthcare accessibility. **[3], [5], [8]**

10. LIMITATIONS

Despite its advantages, the proposed system has certain limitations. The prediction accuracy depends heavily on the quality and size of the training dataset. Incorrect or incomplete patient information may affect the prediction results. The current system is based mainly on clinical symptoms and does not analyze medical images such as CT scans or X-rays. In addition, machine learning models require periodic retraining using updated datasets to maintain high prediction performance. Therefore, the system should be considered as a decision-support tool rather than a replacement for professional medical diagnosis.

[6], [7]

11. FUTURE SCOPE

Several improvements can be incorporated into the proposed Lung Cancer Prediction System in the future. Deep Learning techniques such as Convolutional Neural Networks (CNNs) can be integrated for analyzing CT scan and X-ray images to improve prediction accuracy. Explainable Artificial Intelligence (XAI) can be included to provide interpretable predictions for healthcare professionals. Cloud computing technology can enable secure storage and real-time access to patient data across multiple hospitals. Mobile applications can also be developed to facilitate remote health monitoring and telemedicine services. Integration with Internet of Things (IoT) devices such as smart health sensors can further enhance continuous patient monitoring and early disease detection. **[7], [8], [15]**

12. CONCLUSION

The proposed **Lung Cancer Prediction System** demonstrates the effectiveness of Machine Learning in assisting early diagnosis of lung cancer. By analyzing patient symptoms, demographic information, and lifestyle factors, the system accurately predicts whether a patient is at high or low risk of developing lung cancer. Multiple supervised learning algorithms were implemented and compared using performance metrics such as Accuracy, Precision, Recall, and F1-Score. Among all the evaluated algorithms, Random Forest achieved the highest prediction accuracy of 97%, making it the most suitable classifier for the proposed system. The developed system provides fast, reliable, and cost-effective predictions while supporting healthcare professionals in making informed clinical decisions. Although it cannot replace conventional diagnostic procedures, it serves as an efficient preliminary screening tool capable of reducing diagnostic delays and improving patient outcomes. Future enhancements involving deep learning, cloud computing, medical image analysis, and Explainable AI will further strengthen the capabilities of the proposed healthcare system. **[3], [5], [7], [8]**

REFERENCES

[1] World Health Organization (WHO), *Cancer Fact Sheets*, Geneva, Switzerland, 2024.

[2] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer Statistics, 2024," *CA: A Cancer Journal for Clinicians*, vol. 74, no. 1, pp. 17–48, 2024.

[3] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[4] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[5] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[6] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed. Wiley, 2013.

[7] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[8] K. Kourou et al., "Machine Learning Applications in Cancer Prognosis and Prediction," *Computational and Structural Biotechnology Journal*, vol. 13, pp. 8–17, 2015.

[9] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Morgan Kaufmann, 2012.

[10] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.