Lung Cancer Detection: A Comparative Approach Using Regressor Methods

M. Thangavel
Department of Computer Science, Government Arts College (Autonomous), Salem, India

A.M. Kalpana
Department of Computer Science and Engineering, Government College of Engineering, Dharmapuri, India

S. Ananth
Department of Artificial Intelligence & Data Science, Mahendra Engineering College (Autonomous), Namakkal, India

Abstract. Early and accurate detection of lung cancer is critical due to its high mortality rate and the potential for improved patient outcomes through timely intervention. Traditional diagnostic methods, relying primarily on imaging and clinical assessment, often face challenges such as delayed detection, subjective interpretation, and limited predictive accuracy. This investigation employs a comprehensive lung cancer risk dataset, preprocessed through normalization using StandardScaler and MinMaxScaler, along with feature selection via Recursive Feature Elimination with Support Vector Machines. Class imbalance was addressed using Synthetic Minority Oversampling Technique (SMOTE). Multiple regression and classification algorithms, including Logistic Regression, K-Nearest Neighbors, Gaussian and Multinomial Naive Bayes, Decision Tree, Support Vector Classifier, Random Forest, XGBoost, Gradient Boosting, and Multi-Layer Perceptron, were trained and optimized through RandomizedSearchCV with K-Fold cross-validation. A Voting Classifier combining Random Forest, Decision Tree, and Extra Trees was implemented, and model interpretability was enhanced using explainable AI techniques such as LIME and SHAP, with deployment facilitated via the Flask framework. Evaluation using accuracy, precision, recall, F1-score, and confusion matrices demonstrated that the Voting Classifier outperformed all individual models, achieving 98.1% across all metrics. The integration of ensemble learning with explainable AI provides enhanced diagnostic reliability, transparency, and interpretability, significantly improving lung cancer detection accuracy over conventional approaches.

Keywords: Lung Cancer, Magnetic Resonance imaging, Computed Tomography, Machine Learning, K-fold Cross validation.

1 INTRODUCTION

Lung cancer remains one of the most critical global health challenges due to its high incidence and mortality rates, imposing substantial clinical, social, and economic burdens worldwide [1]. It is consistently ranked among the leading causes of cancer-related deaths, largely because the disease is often diagnosed at an advanced stage when treatment options are limited [2]. Advances in medical imaging, clinical screening, and diagnostic technologies have contributed to improved detection capabilities; however, survival rates remain low compared to other cancer types [3]. Early identification of lung cancer plays a decisive role in improving prognosis, enabling timely intervention, and enhancing overall patient outcomes [4].

Despite ongoing progress, several limitations persist in existing lung cancer detection approaches. Conventional diagnostic practices often rely heavily on clinician expertise and complex imaging interpretation, which may lead to subjective assessments and delayed diagnoses [5]. Furthermore, variability in patient demographics, clinical conditions, and comorbidities introduces additional complexity, reducing the generalizability of current diagnostic frameworks [6]. Many existing systems lack sufficient robustness, interpretability, or consistency when applied across heterogeneous datasets, highlighting a critical gap between technological capability and practical clinical applicability [7].

The objective of this study is to develop a reliable and systematic predictive framework for lung cancer detection that addresses the limitations of existing approaches. By utilizing comprehensive patient risk information and advanced analytical strategies, the proposed work aims to improve diagnostic accuracy, support early-stage detection, and enhance decision-making reliability [8]. The study contributes by emphasizing structured data handling, robust validation, and clinically meaningful outcomes, thereby establishing a foundation for dependable lung cancer screening solutions without increasing diagnostic complexity [9].

The significance of this work lies in its potential to strengthen clinical decision support systems and improve healthcare efficiency through accurate, consistent, and interpretable predictions. By enabling earlier detection and reducing diagnostic uncertainty, the proposed approach can help lower healthcare costs, minimize unnecessary interventions, and improve patient survival rates [10]. Overall, this work seeks to advance the role of data-driven intelligence in lung cancer diagnosis and contribute toward more effective, scalable, and patient-centric healthcare solutions.

2 LITERATURE REVIEW

Recent advancements in machine learning (ML) and deep learning (DL) have significantly influenced the domain of medical diagnostics, particularly in the detection of lung cancer. Sharma et al. [11] proposed a convolutional neural network (CNN) framework integrated with transfer learning to detect lung cancer, achieving improved classification accuracy through feature reuse from pretrained models. Their approach demonstrated strong performance in image-based diagnosis but required substantial computational resources, limiting scalability in low-resource clinical environments. Similarly, Band et al. [12] employed traditional ML algorithms to predict lung cancer using clinical data, presenting a cost-effective alternative to deep architectures. However, their approach faced challenges in handling high-dimensional datasets and lacked model interpretability, highlighting the need for enhanced feature selection and explainability mechanisms.

Sneha et al. [13] introduced a deep learning model for histopathological image classification, showing superior precision in differentiating malignant and benign tissue samples. While their model effectively captured complex spatial patterns, it was constrained by the availability of labeled histopathological data and potential overfitting on small datasets. Khan et al. [14] conducted a systematic review synthesizing multiple ML-based diagnostic methods, emphasizing the variability in model performance due to inconsistent data preprocessing and imbalance handling strategies. The review underscored the importance of unified evaluation frameworks and standardized datasets to ensure generalizable diagnostic outcomes. Devarajan et al. [15] advanced automated lung cancer detection using deep architectures applied to medical imaging data, demonstrating high predictive accuracy. Nevertheless, their study primarily concentrated on image-level classification, with limited exploration of demographic or clinical correlates that could enhance diagnostic reliability.

Indumathi et al. [16] explored the use of ML-based detection and analysis models for lung cancer, focusing on performance evaluation across multiple algorithms. While the study confirmed the potential of data-driven techniques in early detection, it lacked depth in feature optimization and comparative analysis across large datasets. Anbarasi [17] provided a comprehensive review of ML and DL techniques applied to lung cancer prediction, identifying key trends and emerging challenges such as data heterogeneity, model overfitting, and insufficient cross-validation. The review highlighted the necessity of hybridized approaches that balance model complexity and interpretability, motivating further exploration of integrative frameworks. Nakarmi et al. [18] proposed a structured ML framework emphasizing preprocessing and feature engineering for improved detection accuracy. Despite achieving promising results, the study did not adequately address model adaptability across varying patient populations or imaging modalities.

Kumar et al. [19] conducted an extensive survey on exploratory data analysis and ML-based methods for lung cancer detection, stressing the importance of effective data visualization and correlation analysis in model performance. Their work provided valuable insights into the preliminary analytical stages of lung cancer research but fell short of demonstrating a complete predictive modeling pipeline. Bhuiyan et al. [20] investigated the role of predictive modeling in early lung cancer detection, integrating multiple algorithms to enhance public health screening efforts. Although the study presented notable improvements in sensitivity and specificity, it lacked optimization techniques to reduce computational overhead and improve model interpretability.

3 MATERIALS AND METHODS

The proposed system aims to develop an intelligent predictive framework for lung cancer risk assessment using the Lung Cancer Risk Dataset. The workflow encompasses essential data preprocessing, transformation, and model training phases, followed by evaluation and deployment for clinical usability. Multiple machine learning models—including Random Forest, Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree, XGBoost, Gaussian Naive Bayes, Multi-Layer

Perceptron (MLP), and Multinomial Naive Bayes—were employed to establish comparative performance benchmarks. To enhance robustness and accuracy, a Gradient Boosting model and a Voting ensemble combining Random Forest, Decision Tree, and Extra Trees classifiers were implemented. Feature selection and sampling techniques were integrated to address class imbalance and optimize predictive performance. Furthermore, Explainable AI (XAI) methods such as LIME and SHAP were incorporated to provide interpretability of model outputs. A Flask-based user interface enables real-time, transparent risk prediction, promoting clinical accessibility, decision support, and improved patient outcomes.

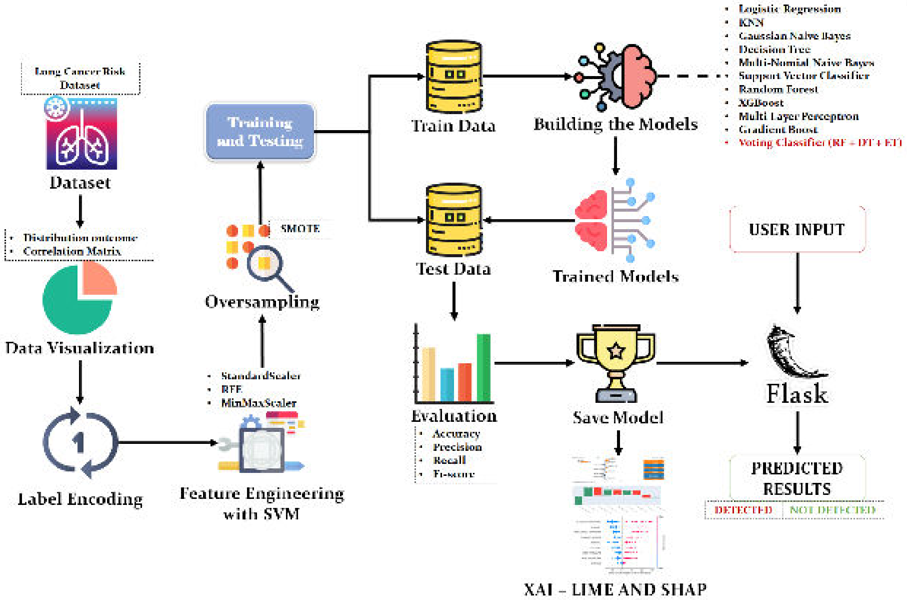


Fig. 1. Proposed Architecture

The system architecture illustrates an end-to-end lung cancer detection pipeline using a risk-based dataset as shown in fig.1. The workflow begins with data visualization and label encoding, followed by feature engineering using SVM-based RFE and normalization with StandardScaler and MinMaxScaler. Class imbalance is handled through SMOTE before training multiple machine learning models. Trained models are evaluated using standard performance metrics, saved, and deployed through a Flask interface, enabling user input–based prediction with explainability supported by LIME and SHAP.

3.1 Dataset Collection:

The Lung Cancer dataset utilized in this study was obtained from publicly accessible repositories, comprising 309 patient records with 16 features each, including demographic, behavioral, and clinical indicators. The dataset contains 14 numerical attributes such as age, smoking habits, fatigue, and alcohol consumption, alongside two categorical attributes representing gender and the target label, LUNG_CANCER, indicating the presence or absence of the disease. Each record provides comprehensive information on common risk factors, symptoms, and lifestyle behaviors associated with lung cancer. Notably, the dataset is well-structured with no missing values, facilitating preprocessing and model training. Despite its moderate size, the dataset captures diverse patient profiles, enabling robust analysis and predictive modeling. Its suitability stems from the inclusion of relevant clinical and behavioral features, making it ideal for early risk assessment and classification tasks compared to generic or less detailed datasets.

	GENDER	AGE	SMOKING	YELLOW_FINGERS	ANXIETY	PEER_PRESSURE	CHRONIC DISEASE	FATIGUE	ALLERGY
0	M	69	1	2	2	1	1	2	1
1	M	74	2	1	1	1	2	2	2
2	F	59	1	1	1	2	1	2	1
3	M	63	2	2	2	1	1	1	1
4	F	63	1	2	1	1	1	1	1

Fig. 2. Lung Cancer Risk Dataset

3.2 Pre-Processing:

The preprocessing phase prepares the lung cancer dataset for effective machine learning by improving data quality, ensuring feature consistency, addressing imbalance, and enabling robust, accurate, and reliable model training.

1. Data Visualization. Data visualization was performed to gain an initial understanding of the dataset and its underlying patterns. Outcome distribution analysis helped identify the proportion of lung cancer and non-cancer cases, revealing potential class imbalance issues. Additionally, a correlation matrix was used to analyze relationships among input features, highlighting dependencies and influential variables. This step is essential for exploratory data analysis, as it supports informed preprocessing decisions, improves feature understanding, and guides the selection of appropriate modeling and balancing strategies.

2. Label Encoding. Label encoding was applied to convert categorical attributes into numerical form, enabling their effective utilization in machine learning models. Since most learning algorithms require numerical inputs, this transformation ensures compatibility while preserving the semantic meaning of categorical variables. Proper encoding maintains consistency across features, prevents misinterpretation of non-numeric data, and allows all patient attributes to contribute meaningfully during model training and evaluation.

3. Normalization. Normalization was carried out to standardize feature values by adjusting them to a common scale with zero mean and unit variance. This process minimizes the influence of features with larger numeric ranges and ensures equal contribution from all attributes. Standardization is particularly important for algorithms sensitive to feature magnitude, as it enhances numerical stability, accelerates convergence, and improves overall predictive performance.

4. Feature Engineering with SVM. Feature engineering was performed using a recursive elimination strategy to identify the most relevant predictors for lung cancer detection. By systematically removing less informative attributes, the model focuses on features with higher discriminative importance. This step reduces dimensionality, improves interpretability, decreases computational complexity, and mitigates overfitting, resulting in a more efficient and reliable learning process.

5. Oversampling. Oversampling was employed to address class imbalance within the dataset by generating additional synthetic samples for the minority class. This technique ensures balanced representation of both cancer and non-cancer cases, reducing model bias toward the majority class. Proper handling of imbalance enhances sensitivity, improves classification fairness, and supports reliable detection of lung cancer cases.

6. Feature Scaling. Min–max scaling was applied to rescale features into a fixed numerical range, ensuring uniformity across all predictors. This step improves numerical consistency and is beneficial for algorithms that are sensitive to feature magnitude. Scaling enhances training stability, supports fair comparison across features, and improves compatibility among multiple classification models used in the system.

3.3 Training and Testing:

The processed dataset was divided into training and testing subsets to enable unbiased evaluation of model performance. This separation ensures that predictive models are assessed using unseen data, providing a realistic measure of generalization capability. Dataset splitting is essential for preventing information leakage, validating reliability, and ensuring robust assessment of the lung cancer detection framework.

3.4 Algorithms:

Random Forest. Random Forest is an ensemble learning algorithm that constructs multiple decision trees and aggregates their predictions to enhance accuracy and generalization. By analyzing diverse subsets of data, it mitigates overfitting and provides robust outcomes across varying conditions. The algorithm evaluates feature importance and captures complex relationships between variables, making it particularly effective in high-dimensional and heterogeneous datasets [21]. Its collective decision-making approach ensures stable predictions, improves performance consistency, and enhances interpretability by revealing key contributing factors in classification tasks.

$$Gini = 1 - \sum_{i=1}^{C} (Pi)^2 \quad (1)$$

Logistic Regression. Logistic Regression is a probabilistic classification method that models the likelihood of a binary outcome based on input features. It establishes a logistic function to estimate the probability of class membership, providing interpretable coefficients that indicate feature influence [22]. The algorithm performs well in scenarios with linear relationships, offering robust and transparent predictions. It enables effective risk stratification, supports feature analysis, and serves as a reliable baseline for comparative evaluation in classification tasks, contributing to accurate and explainable decision-making.

$$\hat{y}_i = \sigma(w^T x + b) = \frac{1}{1 + e^{-(w^T x + b)}} \quad (2)$$

K-Nearest Neighbors (KNN). K-Nearest Neighbors is a non-parametric algorithm that classifies instances based on the majority class of their closest neighbors in the feature space. By measuring similarity among data points, it captures local patterns and adapts to complex distributions without assuming underlying parametric models. The method enhances prediction interpretability by indicating which neighboring instances influenced outcomes. Its ability to handle multi-class classification and nonlinear relationships ensures flexible and intuitive performance, although careful consideration of distance metrics and neighbor selection is essential for optimal results [23].

$$distance(x, X_i) = \sqrt{\sum_{j=1}^{d} \left(x_j - X_{i_j}\right)^2} \quad (3)$$

Support Vector Machine (SVM). Support Vector Machine is a supervised learning algorithm that identifies an optimal hyperplane to separate data points into distinct classes. It maximizes the margin between class boundaries, ensuring robust and accurate classification even in high-dimensional or non-linear feature spaces. By employing kernel functions, SVM captures complex interactions among features and provides confidence estimates for predictions. Its margin-based optimization enhances generalization, while its structured decision boundaries support interpretable and reliable classification, making it well-suited for high-stakes prediction and risk assessment applications [24].

$$minimize \ \frac{1}{2}\,||W||^2 + C\sum_{i=1}^{n} \xi_i \quad (4)$$

Decision Tree. Decision Tree is a hierarchical, rule-based classification algorithm that partitions data according to feature values to form a tree-like structure. Each node represents a decision criterion, while leaves correspond to outcomes, offering clear and interpretable decision paths. The algorithm [25] effectively handles both categorical and numerical variables, manages missing values, and evaluates feature importance. Its transparent structure supports explainable predictions and serves as the foundation for ensemble methods. Decision Trees provide intuitive insights into attribute contributions, enhancing robustness and interpretability in classification tasks.

$$I\,(i) = 1 - \sum_{i=1}^{k} p_i^2 \quad (5)$$

XGBClassifier. XGBClassifier is an optimized gradient boosting algorithm that sequentially builds weak learners to minimize prediction errors and enhance accuracy. It applies regularization to control overfitting and efficiently handles heterogeneous, high-dimensional data. The algorithm combines iterative tree-based models [26], each correcting prior mistakes, to capture complex feature interactions and subtle patterns. Its high computational efficiency, feature importance evaluation, and robust generalization make it suitable for applications requiring precise and reliable predictions, while supporting interpretability in understanding the contributions of individual features.

Gaussian Naïve Bayes (GaussianNB). Gaussian Naïve Bayes is a probabilistic classification algorithm based on Bayes' theorem, assuming normally distributed features. It calculates posterior probabilities for each class, enabling fast and interpretable predictions. The algorithm effectively handles continuous variables, manages uncertainty, and performs well in scenarios with limited data. Its simplicity ensures computational efficiency while providing baseline performance for comparative studies. GaussianNB facilitates understanding of individual feature impacts and contributes to robust probabilistic decision-making, making it a reliable choice for early risk assessment and classification tasks [27].

Multinomial Naïve Bayes (MultinomialNB). Multinomial Naïve Bayes is a probabilistic model designed for discrete feature counts, computing the likelihood of class membership based on observed occurrences. It provides fast and interpretable predictions for categorical or encoded attributes and is robust against high-dimensional data [28]. By estimating feature contributions to class probabilities, it supports explainable decision-making. The algorithm is computationally efficient, offers baseline performance for comparison with more complex models, and contributes to generalizable classification outcomes, particularly in structured datasets with discrete features.

Multi-Layer Perceptron (MLP). Multi-Layer Perceptron is a feedforward neural network that learns complex, non-linear relationships through multiple layers of interconnected nodes. Each layer transforms input features to capture higher-order interactions, enabling improved classification accuracy. The network generalizes well with appropriate regularization and supports high-dimensional, multi-class tasks. Its flexibility allows modeling intricate patterns that linear algorithms cannot capture, while its layer-wise structure facilitates understanding of feature interactions. MLP provides robust, high-performance predictions for tasks requiring sophisticated decision-making [29].

$$\hat{y} = f\left(W^{L} f\left(W^{L-1} \cdots f(W^{1}X + b^{1}) + b^{(L-1)}\right) + b^{L}\right) \quad (6)$$

Gradient Boosting. Gradient Boosting is an ensemble algorithm that iteratively builds weak learners to correct previous prediction errors, improving accuracy and reducing bias. By sequentially optimizing residuals, it captures intricate relationships between input features and output classes. The method includes regularization to prevent overfitting, ensures robust generalization, and handles heterogeneous data efficiently. Its feature importance analysis contributes to interpretability, while its high predictive performance makes it suitable for critical applications requiring reliable classification and precise risk estimation across diverse datasets [30].

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \quad (7)$$

Voting Classifier. Voting Classifier is an ensemble technique that combines multiple base models to produce aggregated predictions through majority voting. By leveraging the complementary strengths of diverse algorithms, it reduces bias and variance while improving stability and overall accuracy. The approach enhances robustness, captures complex feature interactions, and provides more consistent outcomes than individual classifiers. Its transparent structure allows analysis of contributing models, enabling interpretability and reliable decision-making in critical applications requiring balanced and highly accurate predictions.

$$\hat{y} = argmax_c \left(\sum_{i=1}^{n} II\ (\hat{y}_i = c) \right) \quad (8)$$

3.5 Integration of XAI and Flask Framework:

The system incorporates Explainable Artificial Intelligence (XAI) techniques, including LIME and SHAP, to enhance the interpretability of lung cancer predictions. LIME provides local explanations by evaluating the contribution of individual features to a specific prediction, presenting results through interactive visualizations such as waterfall plots. This enables clinicians to understand the influence of risk factors like smoking, fatigue, and coughing on model decisions. SHAP complements this by computing global and local feature attributions, offering a comprehensive overview of feature importance across multiple instances, supporting transparent and trustworthy decision-making in clinical contexts.

The XAI modules are integrated within a Flask-based web framework, allowing real-time prediction explanations for user-provided inputs. The Flask interface serves as a user-friendly portal where input patient data can be analyzed, and corresponding feature contributions are visualized interactively. This combination ensures accessible, interpretable, and clinically relevant insights, bridging advanced machine learning models with practical, explainable, and deployable healthcare applications.

4 EXPERIMENTAL RESULTS

Accuracy: The accuracy of a test is its ability to differentiate the patient and healthy cases correctly. To estimate the accuracy of a test, we should calculate the proportion of true positive and true negative in all evaluated cases. Mathematically, this can be stated as:

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} \quad (9)$$

Precision: Precision evaluates the fraction of correctly classified instances or samples among the ones classified as positives. Thus, the formula to calculate the precision is given by:

$$Precision = \frac{\text{True Positive}}{\text{True Positive } + \text{False Positive}} (10)$$

Recall: Recall is a metric in machine learning that measures the ability of a model to identify all relevant instances of a particular class. It is the ratio of correctly predicted positive observations to the total actual positives, providing insights into a model's completeness in capturing instances of a given class.

$$Recall = \frac{\text{TP}}{\text{TP } + \text{ FN}} (11)$$

F1-Score: F1 score is a machine learning evaluation metric that measures a model's accuracy. It combines the precision and recall scores of a model. The accuracy metric computes how many times a model made a correct prediction across the entire dataset.

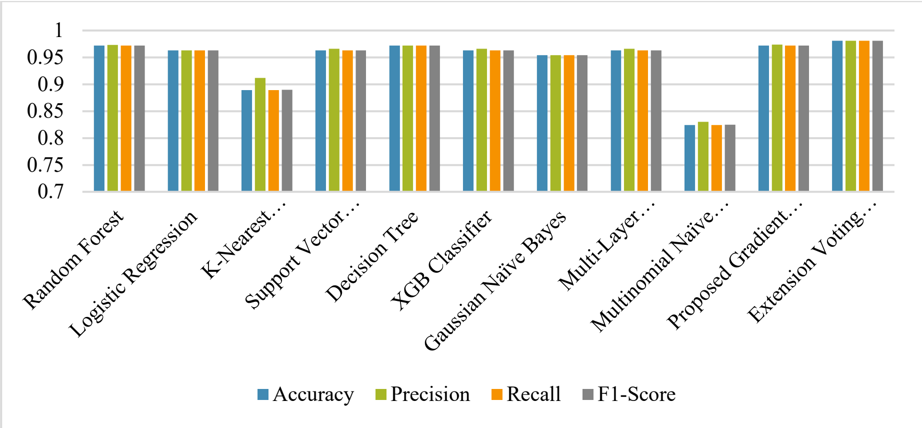
$$F1\ Score = 2 * \frac{Recall \text{ X } Precision}{Recall + Precision} * 100 (12)$$

Table 1. Performance Evaluation Table

ML Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.972	0.973	0.972	0.972
Logistic Regression	0.963	0.963	0.963	0.963
K-Nearest Neighbors (KNN)	0.889	0.912	0.889	0.890
Support Vector Machine (SVM)	0.963	0.966	0.963	0.963
Decision Tree	0.972	0.972	0.972	0.972
XGB Classifier	0.963	0.966	0.963	0.963
Gaussian Naïve Bayes	0.954	0.954	0.954	0.954
Multi-Layer Perceptron (MLP)	0.963	0.966	0.963	0.963
Multinomial Naïve Bayes	0.824	0.830	0.824	0.825
Proposed Gradient Boosting	0.972	0.974	0.972	0.972
Extension Voting (RF + DT + ET)	0.981	0.981	0.981	0.981

In Table.1 compares multiple machine learning models using accuracy, precision, recall, and F1-score, showing that the Extension Voting ensemble outperforms all others with highest values across all metrics.

Fig. 3. Comparison Graph



In Fig.3 visually represents accuracy, precision, recall, and F1-score across all models, highlighting the superior performance of the Extension Voting ensemble, which consistently achieves the highest values among all evaluated algorithms.

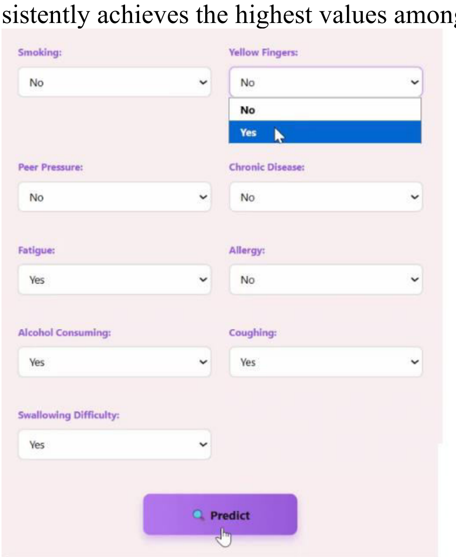


Fig. 4. Upload the Data

In Fig. 4, the input interface allows users to provide patient health and lifestyle information through dropdown menus for lung cancer risk prediction.

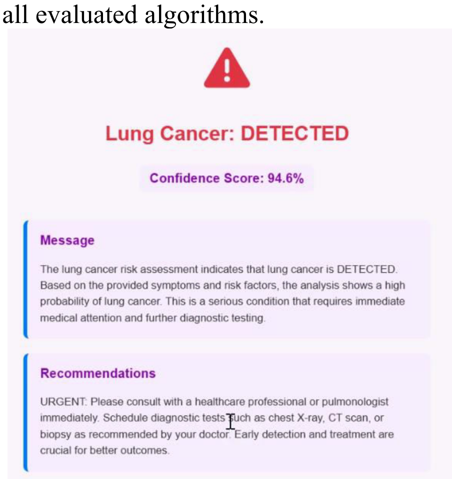


Fig. 5. Predicted Result

In Fig. 5, the output interface displays predicted lung cancer classification with a confidence score of 94.6%, enabling user validation.

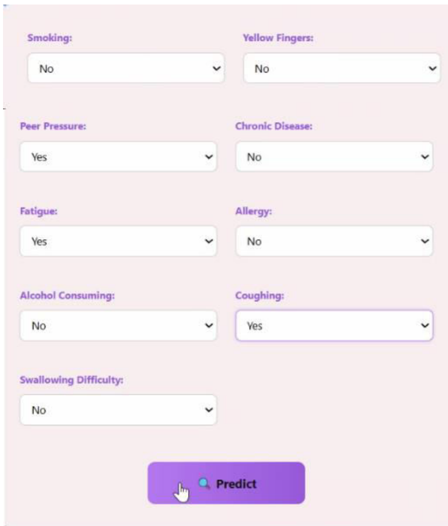


Fig. 6. Upload the Data

In Fig. 6, the input interface allows users to enter patient health details and symptoms for accurate lung cancer risk prediction.

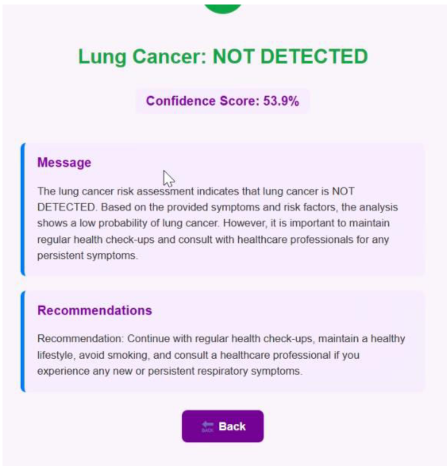


Fig. 7. Predicted Results

In Fig. 7, the output interface shows lung cancer prediction as NOT DETECTED with a confidence score of 53.9% for user reference.

5 CONCLUSION

In conclusion, the primary goal of this system was to improve the accuracy and reliability of lung cancer detection to support early diagnosis and informed clinical decision-making. The approach utilized a Lung Cancer Risk Dataset comprising clinical and demographic attributes, analyzed through a structured machine learning pipeline. Core techniques included data normalization using StandardScaler and MinMaxScaler, feature engineering through Support Vector Machine–based Recursive Feature Elimination, and class imbalance handling using the Synthetic Minority Oversampling Technique. Model training and optimization were performed using RandomizedSearchCV with K-Fold cross-validation across multiple classifiers, including Logistic Regression, K-Nearest Neighbors, Gaussian and Multinomial Naive Bayes, Decision Tree, Support Vector Classifier, Random Forest, XGBoost, Gradient Boosting, and Multi-Layer Perceptron. The effectiveness of the system was validated through performance evaluation using accuracy, precision, recall, and F1-score, where a VotingClassifier combining Random Forest, Decision Tree, and Extra Trees achieved an overall accuracy of 98.1%. To enhance transparency and usability, explainable artificial intelligence techniques such as LIME and SHAP were incorporated, and deployment was facilitated through the Flask framework. Overall, the developed system provides a reliable, interpretable, and automated lung cancer detection solution with strong potential for real-world clinical support and decision-making.

Future work will focus on enhancing the robustness and scalability of lung cancer detection systems by incorporating multimodal medical data, including CT scans, X-rays, and genomic information, to improve diagnostic accuracy. Advanced deep learning architectures, such as hybrid CNN-LSTM models, can be explored for capturing both spatial and temporal patterns in patient data. Additionally, adaptive feature selection techniques and real-time data streaming can be integrated to optimize model performance on large-scale clinical datasets. Research can also investigate federated learning approaches to ensure data privacy while leveraging distributed datasets from multiple healthcare institutions. Ultimately, these enhancements aim to develop a more comprehensive, efficient, and clinically deployable diagnostic framework for early lung cancer detection and risk assessment.

References

1. Jain, V. (2025, April). Lung Cancer Detection using Machine Learning: A Data-Driven Approach. In 2025 8th International Conference on Trends in Electronics and Informatics (ICOEI) (pp. 1345-1350). IEEE.
2. Musthafa, A. S., Santhosh, C., Saran, R., & Suthir, K. (2025, February). Advancing Lung Cancer Diagnosis with Machine Learning: Insights and Innovations. In 2025 4th International Conference on Sentiment Analysis and Deep Learning (ICSADL) (pp. 1300-1307). IEEE.
3. Sharma, R., Dubey, A., Diwakar, M., Gupta, S., Singh, U. P., & Nath, M. K. (2025, April). Lung cancer detection using convolutional neural network and transfer learning. In 2025 3rd International Conference on Advancement in Computation & Computer Technologies (InCACCT) (pp. 773-777). IEEE.
4. Sneha, C., Babu, K. S., Manikanta, K., Manjunadha, V., Kumar, S. D., & Kumar, S. V. P. (2025, February). A Novel Deep Learning Model Based Lung Cancer Detection of Histopathological Images. In 2025 International Conference on Computational, Communication and Information Technology (ICCCIT) (pp. 344-349). IEEE.
5. Band, N. C., Pagrut, M. P., Rawarkar, P. S., Saikhede, P. D., Nage, R. S., & Bahale, A. P. (2024, November). Speculation Of Lung Cancer Using Machine Learning Approach. In 2024 2nd DMIHER International Conference on Artificial Intelligence in Healthcare, Education and Industry (IDICAIEI) (pp. 1-6). IEEE.
6. Musthafa, A. S., Santhosh, C., Saran, R., & Suthir, K. (2025, February). Advancing Lung Cancer Diagnosis with Machine Learning: Insights and Innovations. In 2025 4th International Conference on Sentiment Analysis and Deep Learning (ICSADL) (pp. 1300-1307). IEEE.
7. Choudhury, A. R., Panda, S. R., Mishra, A. K., & Routray, J. (2025, February). Enhancing Lung Cancer Detection by Leveraging Machine Learning Algorithms. In 2025 10th International Conference on Signal Processing and Communication (ICSC) (pp. 369-374). IEEE.
8. Jehangir, B., Nayak, S. R., & Shandilya, S. (2022, January). Lung cancer detection using ensemble of machine learning models. In 2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence) (pp. 411-415). IEEE.
9. Joshi, P. K., Prakash, R., Singh, D., Dhanik, S., Tamta, A., & Sharma, V. K. (2024, November). Lung Cancer Detection: A Machine Learning Approach for Early Diagnosis. In 2024 8th International Conference on Electronics, Communication and Aerospace Technology (ICECA) (pp. 741-746). IEEE.

10. Sebastian, A. M., & Peter, D. (2023). Identifying the predictors from lung cancer data using machine learning. In Sentiment Analysis and Deep Learning: Proceedings of ICSADL 2022 (pp. 691-701). Singapore: Springer Nature Singapore.
11. Sharma, R., Dubey, A., Diwakar, M., Gupta, S., Singh, U. P., & Nath, M. K. (2025, April). Lung cancer detection using convolutional neural network and transfer learning. In 2025 3rd International Conference on Advancement in Computation & Computer Technologies (InCACCT) (pp. 773-777). IEEE.
12. Band, N. C., Pagrut, M. P., Rawarkar, P. S., Saikhede, P. D., Nage, R. S., & Bahale, A. P. (2024, November). Speculation Of Lung Cancer Using Machine Learning Approach. In 2024 2nd DMIHER International Conference on Artificial Intelligence in Healthcare, Education and Industry (IDICAIEI) (pp. 1-6). IEEE.
13. Sneha, C., Babu, K. S., Manikanta, K., Manjunadha, V., Kumar, S. D., & Kumar, S. V. P. (2025, February). A Novel Deep Learning Model Based Lung Cancer Detection of Histopathological Images. In 2025 International Conference on Computational, Communication and Information Technology (ICCCIT) (pp. 344-349). IEEE.
14. Khan, R. H., Miah, J., Nipun, S. A. A., Islam, M., Amin, M. S., & Taluckder, M. S. (2023, September). Enhancing Lung Cancer Diagnosis with Machine Learning Methods and Systematic Review Synthesis. In 2023 8th International Conference on Electrical, Electronics and Information Engineering (ICEEIE) (pp. 1-5). IEEE.
15. Devarajan, H. R., Balasubramanian, S., Swarnkar, S. K., Kumar, P., & Jallepalli, V. R. (2023, December). Deep learning for automated detection of lung cancer from medical imaging data. In 2023 International conference on artificial intelligence for innovations in healthcare industries (ICAIIHI) (Vol. 1, pp. 1-5). IEEE.
16. Indumathi, A., Sathanapriya, M., Vinodh, N., Ashok, M., & Aishwarya, N. (2023, June). Machine Learning based Lung Cancer Detection & Analysis. In 2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS) (pp. 361-365). IEEE.
17. Anbarasi, C. (2025, July). A Comprehensive Review on Lung Cancer Prediction Using ML and DL Techniques. In 2025 6th International Conference on Data Intelligence and Cognitive Informatics (ICDICI) (pp. 1332-1337). IEEE.
18. Nakarmi, A., Sagar, A. K., Musharaf, S., & Jahangir, H. (2022). A Framework for Lung Cancer Detection Using Machine Learning. In Advanced Computing and Intelligent Technologies: Proceedings of ICACIT 2022 (pp. 199-209). Singapore: Springer Nature Singapore.
19. Kumar, K., Ajam, S. M., Kumar, S. H., & Hareshvar, S. P. (2025, February). Comprehensive Survey on Exploratory Data Analysis and Machine Learning Approaches for Lung Cancer Detection. In 2025 3rd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT) (pp. 1791-1797). IEEE.
20. Bhuiyan, M. S., Chowdhury, I. K., Haider, M., Jisan, A. H., Jewel, R. M., Shahid, R., ... & Siddiqua, C. U. (2024). Advancements in early detection of lung cancer in public health: a comprehensive study utilizing machine learning algorithms and predictive models. Journal of Computer Science and Technology Studies, 6(1), 113-121.
21. Joshi, S. S., Bhagat, J., Kulkarni, A., Sharma, S., & Mali, S. M. (2024, July). Lung Cancer Detection and Analysis Using Machine Learning. In International Conference on Applied Artificial Intelligence (pp. 15-20). Cham: Springer Nature Switzerland.
22. Nandula, S. M., Koppisetty, N., Ch, S. R., & Rao, T. R. K. (2023, July). Lung Cancer Detection by Harnessing the Power of Deep Learning with Convolutional Neural Networks. In 2023 2nd International Conference on Edge Computing and Applications (ICECAA) (pp. 1103-1108). IEEE.

23. Babu, L., & Mathiyalagan, P. (2024, August). Diagnosis of Lung Cancer Detection Using Various Deep Learning Models: A Survey. In 2024 5th International Conference on Electronics and Sustainable Communication Systems (ICESC) (pp. 1096-1103). IEEE.
24. Mukhopadhaya, G., Sett, S., Basu, S., Chakraborty, S., Shil, S., Sen, A., & Munyan, P. (2025). Prediction of Lung Cancer Using Supervised Machine Learning. IJSAT-International Journal on Science and Technology, 16(3).
25. Javed, R., Abbas, T., Khan, A. H., Daud, A., Bukhari, A., & Alharbey, R. (2024). Deep learning for lungs cancer detection: a review. Artificial Intelligence Review, 57(8), 197.
26. Mamatha, B., Rashmi, D., Tiwari, K. S., Sikrant, P. A., Jovith, A. A., & Reddy, P. C. S. (2023, August). 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.
27. Bhuvaneshwari, V., Jacklin, J., & Malini, R. (2023, June). Lung Cancer Detection by Artificial Intelligence Using Deep Learning Techniques. In 2023 3rd International Conference on Pervasive Computing and Social Networking (ICPCSN) (pp. 831-839). IEEE.
28. Sonal, D., Pattanaik, S., Nayak, K., Vyas, R. R., Sharma, T., & Kohli, A. (2025, March). An Approach for Detecting and classification of Lung Cancer Using Deep Learning and Machine Learning. In 2025 IEEE 14th International Conference on Communication Systems and Network Technologies (CSNT) (pp. 53-58). IEEE.
29. Shah, S. N. A., & Parveen, R. (2023). An extensive review on lung cancer diagnosis using machine learning techniques on radiological data: state-of-the-art and perspectives. Archives of computational methods in engineering, 30(8), 4917-4930.
30. Agarwal, A. K., Maheshwari, H., Tiwari, R. G., & Jain, V. (2023, December). A Comprehensive Review of Machine Learning Techniques for Lung Cancer Detection. In 2023 12th International Conference on System Modeling & Advancement in Research Trends (SMART) (pp. 623-627). IEEE.