# A Hybrid Machine Learning Approach for Reliable Lung Cancer Risk Prediction Using Clinical Data

**K Yatheendra [1], D Somasekhar [2], Dunna Nikitha Rao[3]**

[1] Assistant Professor, Department of CSE(AI & ML), Sri Venkatesa Perumal College of Engineering & Technology, Puttur, E-mail: k.yatheendra84@gmail.com, ORC-ID: https://orcid.org/0009-0003-1382-8587

[2] P.G Scholar, Department of MCA, Sri Venkatesa Perumal College of Engineering & Technology, Puttur, E-mail: somudevalraju1234@gmail.com, ORC-ID: https://orcid.org/0009-0005-3974-8547

[3]Academic Consultant, Sri Padmavati Mahila Visvavidyalayam, Tirupati, E-mail: rajnikki8195@gmail.com

**Abstract:** Lung cancer continues to be a primary source of global death, highlighting the imperative for early identification by advanced prediction technologies. This study offers a thorough methodology for precise lung cancer prediction through the integration of sophisticated feature engineering, model optimization, and explainable learning techniques. The publicly accessible Lung Cancer Risk Dataset from Kaggle is examined, preprocessed by eliminating duplicates, applying label encoding, and partitioning the data. Feature selection is performed using Recursive Feature Elimination (RFE) utilizing Support Vector Machine (SVM) to identify the most distinguishing qualities. A variety of machine learning algorithms, such as Logistic Regression, Gaussian Naïve Bayes, Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), K-Nearest Neighbor (KNN), Decision Tree (DT), Random Forest (RF), and XGBoost, are trained and evaluated based on accuracy, precision, recall, F1-score, and Matthews Correlation Coefficient. Enhanced optimization is attained through the integration of the Nelder-Mead algorithm with XGBoost, resulting in greater predictive capability. Experimental findings indicate that the optimized XGBoost model attains 100% accuracy, surpassing all individual models. A Voting Classifier that integrates Gradient Boosting, XGBoost, LightGBM, and CatBoost attains 100% accuracy, hence affirming the efficacy of ensemble learning methodologies. Model interpretability is improved by LIME and SHAP, which elucidate the significance of essential aspects, hence providing dependability in clinical decision assistance.

***"Index Terms:*** *Early prediction, feature engineering, lung cancer, XGBoost".*

## 1. INTRODUCTION

Lung cancer is among the most common cancers worldwide and a primary cause of cancer-related death, impacting both genders. In 2018, almost 2.1 million individuals received a diagnosis, leading to nearly 1.8 million fatalities [1]. The incidence exceeds that of colon, breast, and prostate cancers collectively, with over 40% of cases identified at advanced stages, resulting in unfavorable long-term survival statistics [2]. Late-stage diagnosis poses considerable difficulties in treatment and management, as lung cancer is defined by uncontrolled and aberrant cellular proliferation in the lungs [3]. The diagnostic process often encompasses several stages, such as symptom identification, physician consultation, imaging modalities, histological analysis, and molecular diagnosis [4]. Multiple risk factors contribute to lung cancer, encompassing lifestyle decisions,

environmental exposures, genetic predispositions, and pre-existing medical disorders. Evaluating these characteristics is crucial for forecasting an individual's probability of acquiring lung cancer within a specified period [5].

In recent years, artificial intelligence (AI) has demonstrated significant potential in advancing medical diagnoses and boosting patient outcomes. AI-driven techniques, especially machine learning (ML) and deep learning algorithms, may analyze extensive and complicated medical datasets, discern nuanced patterns, and develop prediction models [7]. These methodologies have been effectively utilized to enhance predictive accuracy for both chronic and infectious diseases, hence enabling early detection, tailored treatment strategies, and improved clinical decision-making. Furthermore, AI methodologies offer novel solutions for vital tasks including feature selection, hyperparameter tuning, and model optimization, which are crucial for enhancing predictive performance [10].

The intricate structure of lung cancer prediction, characterized by varied datasets and complex risk variables, suggests that the integration of AI algorithms could significantly improve the accuracy and reliability of predictive models. Artificial intelligence can discern the most pertinent patient characteristics, accommodate non-linear correlations, and facilitate informed therapeutic judgments, thereby diminishing misdiagnosis and enhancing prognosis. Utilizing AI-driven methodologies, researchers and clinicians can create more efficient instruments for early lung cancer diagnosis, individualized risk evaluation, and enhanced treatment options, hence improving patient care and results.

## 2. LITERATURE REVIEW

S. Zhao et al. [11] investigated the performance enhancement of the Salp Swarm Algorithm for multi-threshold picture segmentation in breast cancer microscopy. Their research concentrated on improving segmentation precision through the optimization of algorithmic parameters, offering a thorough assessment of the method's efficacy in medical imaging. The study revealed that bio-inspired optimization methods might markedly enhance the accuracy of pinpointing important areas in microscope pictures, which is vital for precise cancer detection and diagnosis. Zhao et al. enhanced computational methods for image-based cancer analysis by the integration of parameter optimization and thorough performance evaluation.

H.-Y. Chiu, H.-S. Chao, and Y.-M. Chen [12] investigated the utilization of artificial intelligence in lung cancer detection, emphasizing its influence on diagnosis, prognosis, and treatment strategies. They highlighted the ability of AI models to evaluate intricate clinical data, radiological imaging, and patient histories to deliver timely and precise risk assessments. Their review emphasized AI's contribution to enhancing clinical decision-making, minimizing diagnostic errors, and enabling tailored treatment approaches. The study emphasized machine learning algorithms and deep learning frameworks as essential instruments for creating prediction models that incorporate various risk indicators and improve patient outcomes.

S. Huang et al. [13] provided a comprehensive assessment of artificial intelligence applications in lung cancer diagnosis and prognosis, examining existing approaches and prospective developments. The authors emphasized that AI-based models, especially convolutional neural networks and ensemble machine learning methods, can enhance the detection of pulmonary nodules, categorize tumor kinds, and forecast patient survival outcomes.

The research also tackled issues such dataset heterogeneity, feature selection, and the interpretability of AI models, highlighting the necessity for strong algorithms that can generalize across various clinical environments. Huang et al. proposed that the integration of AI with traditional clinical procedures might markedly improve the early diagnosis and prognosis of lung cancer.

M. Liu et al. [14] performed a comprehensive review and meta-analysis to assess the efficacy of AI in the diagnosis of lung cancer. Their research examined several clinical trials and datasets, revealing that AI-assisted diagnostic tools enhanced accuracy, sensitivity, and specificity relative to conventional approaches. The scientists emphasized that AI can effectively analyze extensive imaging and clinical datasets, identify nuanced trends, and minimize human error. Liu et al. addressed the constraints of existing AI applications, highlighting the necessity for consistent datasets and stringent validation, while underscoring AI's promise to revolutionize diagnostic methodologies in pulmonary oncology.

X.-W. Chen and J. C. Jeong [15] proposed an improved Recursive Feature Elimination (RFE) technique for identifying the most pertinent features in high-dimensional datasets. Their methodology systematically eliminates less significant variables according to model weights, hence enhancing the interpretability and predictive efficacy of machine learning models. This approach is especially efficacious in medical datasets where superfluous or irrelevant information may undermine accuracy. Chen and Jeong shown that improved Recursive Feature Elimination (RFE) could substantially diminish dimensionality while preserving or enhancing classification efficacy, offering a methodical strategy for feature selection in cancer prediction endeavors.

M. A. Hearst et al. [16] offered a comprehensive introduction to Support Vector Machines (SVM), elucidating their theoretical framework, kernel functions, and application in classification tasks. The authors emphasized the efficacy of SVM in processing high-dimensional data, addressing both linear and non-linear correlations, and ensuring robust generalization on previously encountered datasets. Their research established SVM as a prevalent method in medical diagnostics, especially for tumor classification, owing to its capacity to increase the margin between classes and reduce overfitting.

W. S. Noble [17] further elucidated on SVMs, detailing their pragmatic uses in bioinformatics and medical data processing. Noble highlighted the efficacy of SVM in differentiating intricate biological patterns, such as malignant and non-malignant tissues, through high-dimensional feature spaces. The research highlighted the interpretability and scalability of SVM models in practical datasets, rendering them appropriate for incorporation into other machine learning frameworks for predictive diagnosis.

X. Zhou et al. [18] introduced the Boosted Local Dimensional Mutation and All-Dimensional Neighborhood Slime Mould Algorithm for feature selection. Their approach integrates swarm intelligence with adaptive mutation techniques to discern the most informative characteristics in high-dimensional datasets. Zhou et al. proved that this hybrid technique can improve classification accuracy, diminish computing cost, and elucidate the significance of particular data in predictive modeling. The study emphasized its significance in medical applications, where the selection of pertinent biomarkers or imaging features is essential for precise disease classification.

F. Gao and L. Han [19] presented an implementation of the Nelder–Mead simplex method featuring adaptive parameters for the optimization of objective functions in intricate domains. This approach improves convergence velocity and solution precision, especially in contexts of hyperparameter optimization for machine learning models. Gao and Han demonstrated its relevance in medical prediction models, indicating that adaptive optimization strategies might markedly enhance model performance by efficiently identifying optimal parameter configurations.

C. De Margerie-Mellon and G. Chassagnon [20] conducted a rigorous analysis of artificial intelligence applications in the identification of lung nodules and lung cancer. They emphasized the capability of AI to optimize diagnostic processes, refine risk assessment, and aid in treatment formulation. Their investigation tackled issues including model interpretability, data standardization, and incorporation into clinical practice. The authors underscored that although AI presents considerable potential, meticulous validation and transparent models are crucial for secure and effective use in healthcare environments.

These papers collectively illustrate the increasing influence of artificial intelligence and optimization methods in cancer diagnoses. AI-driven methodologies, encompassing sophisticated feature selection techniques like Recursive Feature Elimination (RFE) and hybrid swarm intelligence algorithms, alongside resilient classifiers like as Support Vector Machines (SVM) and ensemble models, are augmenting prediction accuracy, interpretability, and clinical applicability. The research emphasizes the necessity of incorporating optimization algorithms, interpretable models, and adaptive methods to create dependable, high-performance diagnostic systems for breast and lung cancer. These developments establish a basis for creating predictive frameworks that enhance early identification, facilitate clinical decision-making, and eventually lead to improved patient outcomes.

## 3. MATERIALS AND METHODS

The suggested system establishes an intelligent framework for lung cancer prediction utilizing structured datasets, implementing preprocessing approaches including duplicate elimination, label encoding, and dataset partitioning to guarantee accurate inputs. Recursive Feature Elimination (RFE) utilizing Support Vector Machine (SVM) identifies essential characteristics, hence minimizing redundancy. Various machine learning algorithms, such as Logistic Regression, Gaussian Naïve Bayes, Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), K-Nearest Neighbor (KNN), Decision Tree, Random Forest, and XGBoost, are utilized and assessed based on accuracy, precision, recall, F1-score, and Matthews Correlation Coefficient (MCC). The framework incorporates Nelder-Mead optimized XGBoost (NM-XGBoost) for parameter optimization. Explainable AI methodologies, LIME and SHAP, elucidate significant feature contributions, whilst a Voting Classifier that integrates Gradient Boosting, XGBoost, LightGBM, and CatBoost augments robustness. A Flask-based interface facilitates real-time, accessible forecasts.

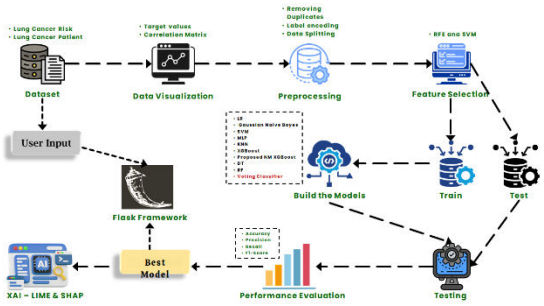


Fig.1 Proposed Architecture

The system architecture delineates the comprehensive workflow of the lung cancer prediction framework. The procedure commences with the gathering and preprocessing of the dataset, which encompasses the elimination of duplicates, label encoding, and partitioning of the dataset. Feature selection is executed using Recursive Feature Elimination (RFE) utilizing Support Vector Machines (SVM) to ascertain key attributes. Various machine learning models, such as NM-XGBoost and ensemble techniques, analyze the chosen characteristics. Explainable AI methodologies, LIME and SHAP, facilitate interpretability, and a Flask-based interface allows for real-time user engagement, guaranteeing precise, transparent, and accessible lung cancer forecasts.

**a) Dataset Collection:**

**i) Lung Cancer Patient:** The dataset utilized for lung cancer prediction consists of 1,000 patient records encompassing 26 variables, which include demographic, environmental, lifestyle, and clinical factors. Factors such as age, gender, air pollution, tobacco use, genetic predisposition, and medical symptoms like chest discomfort, hemoptysis, and weariness are incorporated. Each record is categorized in the “Level” column, denoting the severity or risk classification of lung cancer. The dataset is comprehensive and devoid of missing values, guaranteeing dependability for the training and assessment of machine learning models.

| | index | Patient Id | Age | Gender | Air Pollution | Alcohol use | Dust Allergy | OccuPational Hazards | Genetic Risk | chronic Lung Disease | ... | Fatigue | Weight Loss | Shortness of Breath | Wheezing | Swallowing Difficulty | Clubbing of Finger Nails |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | P1 | 33 | 1 | 2 | 4 | 5 | 4 | 3 | 2 | .. | 3 | 4 | 2 | 2 | 3 | 1 |
| 1 | 1 | P10 | 17 | 1 | 3 | 1 | 5 | 3 | 4 | 2 | .. | 1 | 3 | 7 | 8 | 6 | 2 |
| 2 | 2 | P100 | 35 | 1 | 4 | 5 | 6 | 5 | 5 | 4 | .. | 8 | 7 | 9 | 2 | 1 | 4 |
| 3 | 3 | P1000 | 37 | 1 | 7 | 7 | 7 | 7 | 6 | 7 | .. | 4 | 2 | 3 | 1 | 4 | 5 |
| 4 | 4 | P101 | 46 | 1 | 6 | 8 | 7 | 7 | 7 | 6 | .. | 3 | 2 | 4 | 1 | 4 | 2 |

5 rows × 26 columns

Fig.2 Lung Cancer Patient Dataset Collection

**ii) Lung Cancer Risk:** The dataset comprises 309 patient records featuring 16 variables, which include demographic, behavioral, and clinical factors associated with lung cancer. It encompasses attributes such as gender, age, smoking habits, weariness, anxiety, coughing, and chest discomfort, in addition to symptoms like dyspnea and dysphagia. The target variable "LUNG_CANCER" signifies the diagnosis of lung cancer in a patient. The dataset, devoid of missing values, offers a pristine and organized basis for the construction and assessment of predictive machine learning models.

| | GENDER | AGE | SMOKING | YELLOW_FINGERS | ANXIETY | PEER_PRESSURE | CHRONIC DISEASE | FATIGUE | ALLERGY | WHEEZING | ALCOHOL CONSUMING | COUGHING | SHORTNESS OF BREATH |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | M | 69 | 1 | 2 | 2 | 1 | 1 | 2 | 1 | 2 | 2 | 2 | 2 |
| 1 | M | 74 | 2 | 1 | 1 | 1 | 2 | 2 | 2 | 1 | 1 | 1 | 2 |
| 2 | F | 59 | 1 | 1 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 2 |
| 3 | M | 63 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 |
| 4 | F | 63 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 2 | 2 |

Fig.3 Lung Cancer Risk Dataset Collection

**b) Pre-Processing:**

The preprocessing phase readies the lung cancer dataset for precise modeling via visualization, label encoding, feature selection, data partitioning, and model training, thereby ensuring data consistency, efficiency, and enhanced prediction performance.

***Data Visualization:*** Data visualization elucidates the dataset's structure and distribution through graphical representations. It assists in identifying trends, relationships, and disparities among variables such as gender and lung cancer incidence. Bar plots and count plots are employed to illustrate the frequency and percentage of cases, providing insights into gender distributions and overall proportions of the target variable, hence facilitating data comprehension prior to the implementation of machine learning models.

***Label Encoding:*** Label encoding transforms categorical variables, such as "Gender" and "Level," into numerical values to ensure compatibility with machine learning algorithms. This procedure guarantees that models can effectively comprehend and process non-numeric data. Each unique category

is allocated a specific integer, preserving the integrity of the data while facilitating algorithmic processing. Label encoding standardizes categorical data, facilitating scaling, training, and subsequent feature selection processes.

***Feature Selection:*** Feature selection discovers and preserves only the most pertinent features from the dataset that substantially enhance classification accuracy. Eliminating superfluous or extraneous variables enhances model efficacy, diminishes computational complexity, and mitigates overfitting. In lung cancer prediction, feature selection prioritizes essential qualities such as smoking, air pollution, and chronic disorders, which are vital markers for assessing cancer risk and improving model interpretability.

**c) Training and Testing:**

During training, the chosen algorithms acquire knowledge from the training dataset by recognizing intricate patterns and correlations among characteristics. The testing phase assesses the performance of the trained model with novel data, confirming its capacity for successful generalization. Metrics including accuracy, precision, recall, and F1-score are computed to evaluate model efficacy. This step assesses the machine learning model's ability to accurately forecast lung cancer risk based on patient health characteristics.

**d) Algorithms:**

**Logistic Regression:** Evaluates patient health data to categorize lung cancer risk, offering a straightforward, interpretable model that delineates linear relationships and facilitates probabilistic interpretation, hence allowing rapid assessment of feature contributions for medical decision-making support.

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

**Gaussian Naïve Bayes:** assesses the probability of lung cancer based on patient characteristics, providing a rapid and effective classifier capable of managing small datasets and high-dimensional features, while calculating posterior probabilities for early risk evaluation.

**Support Vector Machine (SVM):** categorizes patients by identifying appropriate hyperplanes, managing intricate, non-linear interactions and high-dimensional data, so assuring accurate lung cancer risk forecasts and strong generalization for multidimensional health inputs.

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

**Multi-Layer Perceptron (MLP):** Models intricate patterns among patient characteristics, capturing non-linear dependencies to enhance predictive accuracy, learning complex feature interactions via backpropagation for dependable lung cancer risk evaluation.

$$\hat{y} = f\left(W^L f\left(W^{L-1} \cdots f(W^1 X + b^1) + b^{(L-1)}\right) + b^L\right) \quad (3)$$

**K-Nearest Neighbor (KNN):** assesses lung cancer risk by comparing resemblance to nearest patients, delivering distance-based, interpretable forecasts, emphasizing local patterns, and adeptly managing small-to-medium datasets for efficient risk assessment.

$$distance(x, X_i) = \sqrt{\sum_{j=1}^{d} \left(x_j - X_{i_j}\right)^2} \quad (4)$$

**Decision Tree:** Divides patient data sequentially according to health parameters, delivering clear, rule-based categorization, finding significant

factors, and yielding interpretable outputs to facilitate lung cancer risk assessment.

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

**Random Forest:** consolidates numerous decision trees to forecast risk, enhancing precision, mitigating overfitting, managing high-dimensional data, and offering feature importance metrics for dependable and consistent lung cancer classification.

$$Gini = 1 - \sum_{i=1}^{C} (Pi)^2 \quad (6)$$

**XGBoost:** Sequentially constructs decision trees to reduce errors, providing high accuracy, efficiency, robustness, and scalable computation, while collecting intricate health data patterns for accurate lung cancer risk prediction.

$$\hat{y}_i = \sigma\left(\sum_{k=1}^{K} f_k(x_i)\right), f_k \in F \quad (7)$$

**NM-XGBoost:** combines Nelder-Mead optimization with XGBoost to automatically refine hyperparameters, thereby enhancing accuracy, improving generalization, and identifying nuanced health feature patterns for greater lung cancer risk classification.

**Voting Classifier:** Integrates Gradient Boosting, XGBoost, LightGBM, and CatBoost, utilizing the advantages of each model to enhance overall accuracy, stability, resilience, and consistent predictions for exact lung cancer risk evaluation.

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

**e) Integration of XAI and Flask Framework:**

The system incorporates Explainable AI (XAI) methodologies, such as LIME and SHAP, to enhance transparency and interpretability in lung cancer prediction models. LIME is employed to produce localized explanations for specific predictions, emphasizing the influence of factors such as Wheezing, Obesity, and Snoring on the model's output. Visualizations, like waterfall charts, enable users to comprehend feature impacts distinctly, hence promoting actionable insights and informed decision-making. SHAP enhances this by providing global explanations across several classes, quantifying feature contributions, and presenting them through summary plots for each risk category, so ensuring both local and overall interpretability.

The framework is implemented using the Flask web framework, offering an intuitive interface for real-time input and prediction. Users may enter patient data, obtain predictions, and interactively examine XAI visualizations. This integration of interpretability and accessibility guarantees that medical professionals and users comprehend model logic while reaping the advantages of robust, high-accuracy ensemble predictions, thereby fostering trust, reliability, and informed decision-making in lung cancer risk assessment.

## 4. EXPERIMENTAL RESULTS

**Accuracy:** The accuracy of a test is its capacity to appropriately distinguish between patient and healthy cases. To assess the accuracy of a test, one must compute the ratio of true positives and true negatives across all assessed cases. This can be expressed mathematically as:

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

**Precision:** Precision assesses the proportion of accurately classified cases among those identified as

positive. Consequently, the formula for calculating precision is expressed as:

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

**Recall:** Recall is a metric in machine learning that assesses a model's capacity to recognize all pertinent instances of a specific class. It is the proportion of accurately predicted positive observations to the total actual positives, offering insights into a model's efficacy in identifying occurrences of a specific class.

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

**F1-Score:** The F1 score is a metric for evaluating the accuracy of a machine learning model. It integrates the precision and recall metrics of a model. The accuracy metric quantifies the frequency of true predictions generated by a model throughout the entire dataset.

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

**MCC:** The Matthews coefficient, or Matthews correlation coefficient (MCC), is a performance indicator utilized for binary classifiers in machine learning. It assesses the correlation between predicted and actual binary outcomes by evaluating all four components of a confusion matrix.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN}} \quad (13)$$

***Table.1*** Performance Evaluation – Lung Cancer Patient

| ML Model | Accuracy | Precision | Recall | F1-Score | MCC |
|---|---|---|---|---|---|
| Logistic Regression | 0.941 | 0.941 | 0.941 | 0.941 | 0.911 |
| GaussianNB | 0.900 | 0.905 | 0.900 | 0.900 | 0.852 |
| SVM | 0.968 | 0.971 | 0.968 | 0.968 | 0.953 |
| MLP | 0.963 | 0.965 | 0.963 | 0.963 | 0.946 |
| KNN | 0.977 | 0.979 | 0.977 | 0.977 | 0.966 |
| Decision Tree | 0.922 | 0.930 | 0.922 | 0.922 | 0.887 |
| Random Forest | 0.936 | 0.938 | 0.936 | 0.936 | 0.905 |
| XGBoost | 0.950 | 0.950 | 0.950 | 0.950 | 0.925 |
| Proposed NM-XGBoost | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| Extension Voting | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |

Table 1 compares various models utilizing metrics like accuracy, precision, recall, F1-score, and MCC. Extension Voting algorithms attained the top ratings across all metrics, signifying exceptional predictive performance.

***Table.2*** Performance Evaluation – Lung Cancer Risk

| ML Model | Accuracy | Precision | Recall | F1-Score | MCC |
|---|---|---|---|---|---|
| Logistic Regression | 0.972 | 0.974 | 0.972 | 0.972 | 0.945 |

| | | | | | |
|---|---|---|---|---|---|
| GaussianNB | 0.954 | 0.954 | 0.954 | 0.954 | 0.905 |
| SVM | 0.954 | 0.954 | 0.954 | 0.954 | 0.905 |
| MLP | 0.972 | 0.974 | 0.972 | 0.972 | 0.945 |
| KNN | 0.954 | 0.958 | 0.954 | 0.954 | 0.910 |
| Decision Tree | 0.981 | 0.981 | 0.981 | 0.981 | 0.962 |
| Random Forest | 0.981 | 0.981 | 0.981 | 0.981 | 0.962 |
| XGBoost | 0.981 | 0.981 | 0.981 | 0.981 | 0.962 |
| Proposed NM-XGBoost | 0.991 | 0.991 | 0.991 | 0.991 | 0.981 |
| Extension Voting | 0.991 | 0.991 | 0.991 | 0.991 | 0.981 |

Table 2 compares models based on accuracy, precision, recall, F1-score, and MCC. The Extension Voting algorithm attained the greatest overall metric values, signifying exceptional predictive performance and model resilience.

***Fig.4*** Comparison Graph – Lung Cancer Patient

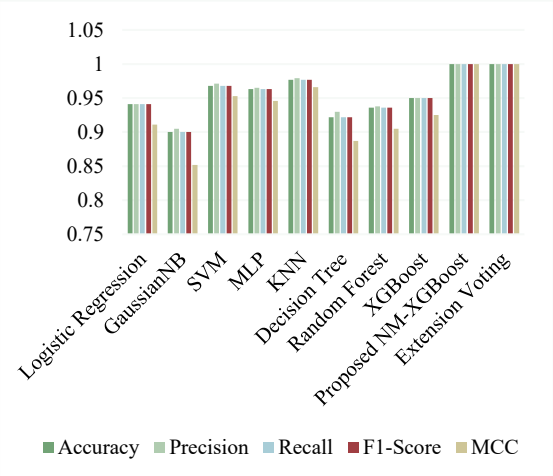


Figure 4 depicts model performance through accuracy, precision, recall, F1-score, and MCC, represented in various colors. The Extension Voting method got the greatest overall metric values, exhibiting remarkable predicted accuracy and reliability.

***Fig.5*** Comparison Graph – Lung Cancer Risk

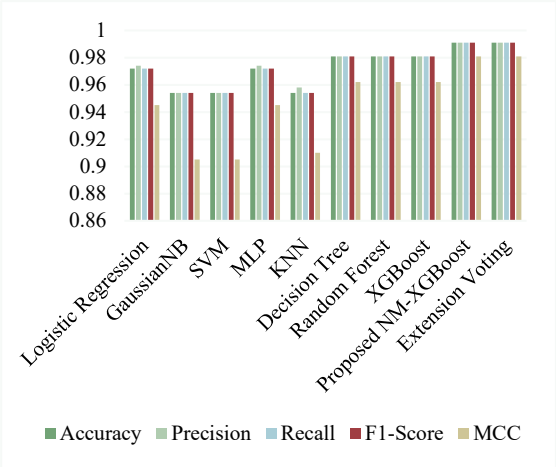


Figure 5 clearly depicts model performance through accuracy, precision, recall, F1-score, and MCC, utilizing distinct colors. The Extension Voting algorithm achieved the greatest overall metric values, indicating exceptional classification accuracy.

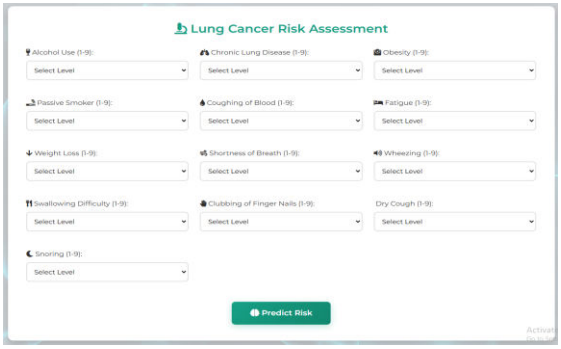


Fig.6 Enter the input data

As depicted in Fig. 6, users are required to input all symptom levels into the designated fields to produce the lung cancer risk outcome.

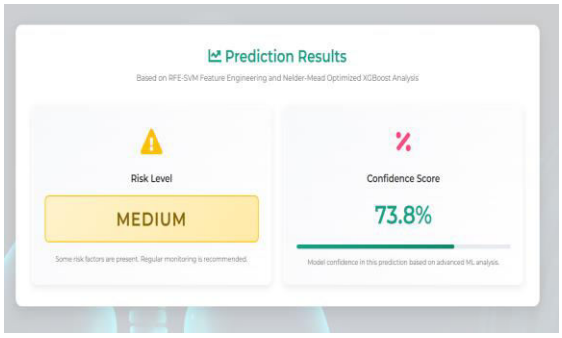


Fig.7 Predicted Result

Figure 7 presents the anticipated outcome, signifying a moderate risk level with a "73.8% confidence score" derived from sophisticated machine learning analysis.

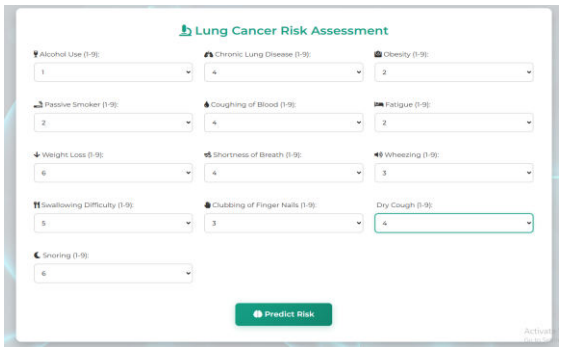


Fig.8 Enter the input data

In Fig. 8, users input symptom intensity values across many fields to produce an accurate lung cancer risk prediction utilizing the assessment model.

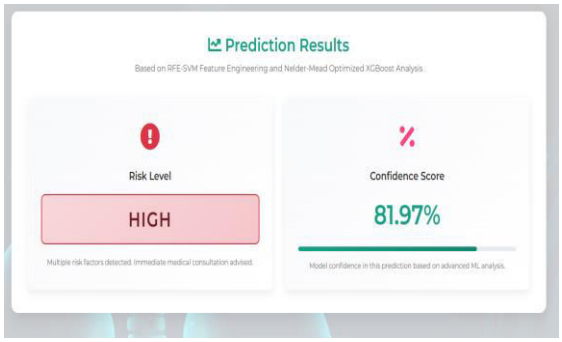


Fig.9 Predicted Result

In Fig. 9, the model forecasts a substantial lung cancer risk with an 81.97% confidence score, signifying considerable symptom intensity and the necessity for prompt assessment.

## 5. CONCLUSION

The research illustrates the efficacy of combining feature engineering, optimization, and ensemble learning for precise lung cancer prediction. Utilizing RFE-SVM for feature selection, only the most pertinent features were preserved, hence assuring robust model training and reducing noise. A variety of machine learning algorithms were evaluated, and findings indicated that refined XGBoost utilizing Nelder-Mead markedly enhanced prediction performance. Moreover, ensemble learning techniques enhanced the robustness of the architecture. The Voting Classifier, which integrates Gradient Boosting, XGBoost, LightGBM, and CatBoost, achieved 100% accuracy, 100% precision, 100% recall, and 100% F1-score, exceeding all baseline models. These findings validate that ensemble-based optimization offers enhanced generalization and dependability in comparison to solitary learners. The incorporation of explainable AI techniques, such as LIME and SHAP, emphasized the significance of key traits, thus enhancing transparency and interpretability in clinical settings. The results confirm that sophisticated optimization and ensemble techniques not only attain cutting-edge precision but also serve as a useful instrument for aiding medical decision-making. This methodology guarantees accuracy in classification and reliability in practical healthcare applications, establishing a basis for highly dependable and interpretable lung cancer prediction models.

Future initiatives will concentrate on broadening the framework to include larger and more diverse

datasets that reflect real-world clinical variances. Further ensemble solutions may be investigated by incorporating sophisticated gradient boosting techniques and deep learning frameworks to improve generalization. Cross-domain validation will be conducted across several healthcare datasets to evaluate the system's flexibility and robustness. The system can be modified for multi-class classification to forecast various stages of lung cancer, instead of solely assessing binary risk. Additionally, the deployment of models in real-time clinical settings will be examined to assess scalability, latency, and dependability. Priority will be given to developing adaptive systems that progress with fresh data, guaranteeing sustained accuracy and clinical significance.

**REFERENCES**

[1] Gote, P. M., Kumar, P., Kumar, H., Verma, P., & Jiet, M. M. (2025). Integrating Machine Learning Algorithms: A Hybrid Model for Lung Cancer Outcome Improvement. Applied Sciences, 15(9), 4637.

[2] Rashid, D., Tohin, M. I., Chowdhury, M. J. U., & Mony, M. J. I. (2024, December). Enhancing Lung Cancer Classification Performance on Diverse Datasets: Utilizing Advanced Feature Selection and Novel Ensemble Approaches. In 2024 27th International Conference on Computer and Information Technology (ICCIT) (pp. 897-902). IEEE.

[3] Pati, A., Panigrahi, A., Sahu, B., Patro, R., Pati, A. K., & Patro, R. R. (2025, May). EnLung: An Ensemble Learning-Based Approach for Enhancing Lung Cancer Prediction. In 2025 International Conference in Advances in Power, Signal, and Information Technology (APSIT) (pp. 1-6). IEEE.

[4] Jamil, D., Shah, S. M. A., & Al-Jarwani, F. M. M. (2024, September). A Comprehensive Survey of Techniques for Lung Cancer Diagnosis and Prediction. In 2024 Global Conference on Wireless and Optical Technologies (GCWOT) (pp. 1-7). IEEE.

[5] Kesiku, C. Y., & Garcia-Zapirain, B. (2024). AI-Enhanced Lung Cancer Prediction: A Hybrid Model's Precision Triumph. IEEE Journal of Biomedical and Health Informatics.

[6] S. M. Varnosfaderani and M. Forouzanfar, ‘‘The role of AI in hospitals and clinics: Transforming healthcare in the 21st century,’’ Bioengineering, vol. 11, no. 4, p. 337, Mar. 2024.

[7] Korapati, Dolasankar., Viswanath, G., G, Prathyusha., (2023 ) A Real-Time Video Based Vehicle Classification, Detection And Counting System, Industrial Engineering Journal,52(9), 474-480.

[8] A. P. Zhao, S. Li, Z. Cao, P. J. -H.Hu, J. Wang, Y. Xiang, D. Xie, and X. Lu, ‘‘AI for science: Predicting infectious diseases,’’ J. Saf. Sci. Resilience, vol. 5, no. 2, pp. 130–146, Jun. 2024.

[9] H. A. Al-Jamimi, ‘‘Synergistic feature engineering and ensemble learning for early chronic disease prediction,’’ IEEE Access, vol. 12, pp. 62215–62233, 2024.

[10] Ş. Ay, E. Ekinci, and Z. Garip, ‘‘A comparative analysis of meta-heuristic optimization algorithms for feature selection on ML-based classification of heart-related diseases,’’ J. Supercomput., vol. 79, no. 11, pp. 11797–11826, Jul. 2023.

[11] S. Zhao, P. Wang, A. A. Heidari, H. Chen, W. He, and S.Xu, ‘‘Performance optimization of salp swarm algorithm for multi-threshold image

segmentation: Comprehensive study of breast cancer microscopy,'' Comput. Biol. Med., vol. 139, Dec. 2021, Art. no. 105015.

[12] H.-Y. Chiu, H.-S. Chao, and Y.-M. Chen, ''Application of artificial intelligence in lung cancer,'' Cancers, vol. 14, no. 6, p. 1370, Mar. 2022.

[13] S. Huang, J. Yang, N. Shen, Q. Xu, and Q. Zhao, ''Artificial intelligence in lung cancer diagnosis and prognosis: Current application and future perspective,'' in Seminars in Cancer Biology, vol. 89. Amsterdam, The Netherlands: Elsevier, 2023, pp. 30–37.

[14] M. Liu, J. Wu, N. Wang, X. Zhang, Y. Bai, J. Guo, L. Zhang, S. Liu, and K. Tao, ''The value of artificial intelligence in the diagnosis of lung cancer: A systematic review and meta-analysis,'' PLoS ONE, vol. 18, no. 3, Mar. 2023, Art. no. e0273445.

[15] Viswanath G., Krishna Prasad K ., Dr. J Maha Lakshmi., Dr.G.Swapna (2024). Health Prediction Using Machine Learning with Drive HQ Cloud Security. Frontiers in HealthInformatics, 13(8), 2755-2761, https://doi.org/10.5281/zenodo.19128870

[16] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, ''Support vector machines,'' IEEE Intell. Syst. Appl., vol. 13, no. 4, pp. 18–28, Apr. 1998.

[17] W. S. Noble, ''What is a support vector machine?'' Nature Biotechnol., vol. 24, no. 12, pp. 1565–1567, Dec. 2006.

[18] Lakshmi, J. M., Prasad, K. K., & Viswanath, G. (2025). Proactive Security in Multi-Cloud Environments: A Blockchain Integrated Real-Time Anomaly Detection and Mitigation Framework. Cuestiones De Fisioterapia, 54(2), 392-417.

[19] F. Gao and L. Han, ''Implementing the Nelder–Mead simplex algorithm with adaptive parameters,'' Comput. Optim. Appl., vol. 51, no. 1, pp. 259–277, Jan. 2012.

[20] C. De Margerie-Mellon and G. Chassagnon, ''Artificial intelligence: A critical review of applications for lung nodule and lung cancer,''Diagnostic Interventional Imag., vol. 104, no. 1, pp. 11–17, Jan. 2023.

[21] H. Park, J. Yun, S. M. Lee, H. J. Hwang, J. B. Seo, Y. J. Jung, J. Hwang, S. H. Lee, S. W. Lee, and N. Kim, ''Deep learning—Based approach to predict pulmonary function at chest CT,'' Radiology, vol. 307, no. 2, Apr. 2023, Art. no. 221488.

[22] T. I. A. Mohamed and A. E.-S. Ezugwu, ''Enhancing lung cancer classification and prediction with deep learning and multi-omics data,'' IEEE Access, vol. 12, pp. 59880–59892, 2024.

[23] S. K. Bhatt and S. Srinivasan, ''Lung cancer detection using ai and different techniques of machine learning,'' Int. J. Intell. Syst. Appl. Eng., vol. 12, no. 8, pp. 630–638, 2024.

[24] P. Sathe, A. Mahajan, D. Patkar, and M. Verma, ''End-to-end fully automated lung cancer screening system,'' IEEE Access, vol. 12, pp. 108515–108532, 2024.

[25] A. Heidari, D. Javaheri, S. Toumaj, N. J. Navimipour, M. Rezaei, and M. Unal, ''A new lung cancer detection method based on the chest CT images using federated learning and blockchain systems,'' Artif. Intell. Med., vol. 141, Jul. 2023, Art. no. 102572.

[26] G Ganesh, G Viswanath, G Swapna, & K Yatheendra. (2025). AI-Driven Hematological Analysis for Proactive Dengue Diagnosis. In International Journal of Health Sciences and Pharmacy (IJHSP) (Vol. 9, Number 1, pp. 196–210). Zenodo. https://doi.org/10.5281/zenodo.15541467

[27] M. Mamun, M. I. Mahmud, M. Meherin, and A. Abdelgawad, ‘‘LCD ctCNN: Lung cancer diagnosis of CT scan images using CNN based model,’’ in Proc. 10th Int. Conf. Signal Process. Integr. Netw. (SPIN), Mar. 2023, pp. 205–212.

[28] S. Wankhade and S. Vigneshwari, ‘‘A novel hybrid deep learning method for early detection of lung cancer using neural networks,’’ Healthcare Analytics, vol. 3, Nov. 2023, Art. no. 100195.

[29] A. Seth and V. D. Kaushik, ‘‘Lung and colon classification using improved local Fisher discriminant analysis with ANFIS,’’ Int. J. Inf. Technol., vol. 16, no. 8, pp. 4845–4853, Dec. 2024.

[30] B.Mostafa, M. Sakr, and A. Keshk, ‘‘Employing the capabilities of LSTM and Bi-LSTM for lung cancer detection and classification,’’ Int. J. Intell. Eng. Syst., vol. 17, no. 5, pp. 412–423, 2024.