International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0456
e ISSN: 2584-2854
Volume: 03
Issue: 09 September 2025
Page No: 2891 - 2896

# Optimizing Lung Cancer Diagnosis with ML Classification Methods

*Meena Preethi[1], Seethalakshmy Anantharaman[2], Lekha[3], Umadevi[4], Paramasivam[5], Hari Shankar[6]*
*[1]Associate Professor, Dept. Of Software System and AIML, Sri Krishna arts and science college, Coimbatore, Tamil Nadu, India.*
*[2]Associate Professor, Dept. Of Psychology, Rathinam College of Arts and Science, Coimbatore, Tamil Nadu, India.*
*[3]Associate Professor, School of sciences, Christ University, India.*
*[4,5,6]PG Scholar - Dept. Of SS, Sri Krishna arts and science college, Coimbatore, Tamil Nadu, India.*
***Emails:*** *meenapreethib@skasc.ac.in[1], hod.psy@rathinam.in[2], lekha.j@christuniversity.in[3], umaramakrishnan5522@gmail.com[4], paramasivam995253@gmail.com[5], lbhari007shankar@gmail.com[6]*

## Abstract
*One of the most advanced causes of deaths due to cancer that occurs all over the world is lung cancer since such cancer is not easily diagnosed in early stages and it takes advance staging so that the disease can be diagnosed. This research study is proposed to learn how machine learning system assists the artificial intelligence to foresee the invasion of lung cancer by early staging through supporting the symptoms and the lifestyle data that have been reported by the patients. A well-organized record of clinical and behavioral features was a preprocessed dataset to ensure that the model was ready through feature encoding and normalization. To categorize the incidence of lung cancer, they have employed some of the supervised learning algorithms such as logistic regression (LR), decision trees (DT) and ensemble strategies. AdaBoost model showed better results than the others and was therefore used in classification of the three types of lung cancer, i.e. the three namely, the small cell carcinoma, adenocarcinoma and the large cell carcinoma. As depicted in the comparison research, it was uncovered how ML could be utilized as a beneficial diagnostic procedure to ascertain the risk of lung cancer and thus be capable of prompt and more informed medical treatment. This study reflects the significance of data-driven practices in completing the existing diagnostic systems and aiding the clinical decision-making in oncology.*

***Keywords:*** *Lung cancer, Artificial intelligence, Supervised learning, Machine learning.*

## 1. Introduction
Lung cancer has a fatality rate and is considered one of the most widespread and most dangerous diseases on the globe due to its fast development and tendency to be identified late. A huge number of cancer mortalities are because of lung cancer proving that early diagnosis and effective medication are of extreme importance [1], [3]. Though these methods are helpful, conventional ways of diagnosis such as biopsies, MRI, CT scans and chest X-rays have serious limitations. These techniques are invasive, costly, time-consuming and not readily available. Therefore, because of this, there is an increased need to come up with alternative methods that can aid in the early detection of the risk of lung cancer. Supervised ML algorithms have been found in use in analyzing large sets of data and discovering implicit patterns results and making predictive decision-making using input features in, [2], [6]. This has created new avenues for developing intelligent systems which can aid clinical decisions and disease detection. Using ML, we were able to improve the rules that are used in scoring. On the example of patient survey data, which also includes demographic characteristics, lifestyle-specific results, such as smoking and alcohol consumption habits, and clinical symptoms, including coughing and chest pain, it can be stated that predictive models can be constructed that will estimate the possibility of lung cancer in patients [3], [5]. The target of this effort is the development of a robust ML design capable of

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0456
e ISSN: 2584-2854
Volume: 03
Issue: 09 September 2025
Page No: 2891 - 2896

assisting in the early detection of lung cancer on the foundation of structured survey responses. The study employs a four-step approach, starting with extensive data collection and pre-processing, encoding and normalization of different features. Various supervised classification algorithms are built and compared to each other to identify the best one applicable to this two-class classification problem [6], [4], [5]. This comparison examines how well each model was able to categorize people as lung cancer- positive or negative lung cancer and this rule will assist doctors in prioritizing high-risk individuals to further investigation.

## 2. Literature Survey

Abdulrahman Alzahrani [1] introduced an enhanced lung cancer detection model that combines the RF algorithm with Conditional Tabular CTGAN to generate synthetic data. The proposed approach demonstrated significant improvements in predictive performance and model reliability. Various classification algorithms, including SVM, KNN, and DT, were used for comparative analysis to assess the model's consistency. The findings revealed that the CTGAN-RF framework outperformed conventional methods, especially in managing class imbalance and improving detection accuracy. Wayahdi and Ruziq examined, regarding lung cancer, the range of machine learning techniques, namely, KNN and XG Boost, as a means of making predictions [4]. The early and reliable prevention is crucial in enhancing the treatment outcome, because lung cancer is a health issue of concern globally. Among their tests was the evaluation of the performance of both models regarding their predictive powers. The test revealed that both KNN and XG Boost yielded good results, but XG Boost was more efficient because it was capable of better description of latent patterns within the data. Hu et al [6]. suggested an effective and scalable method of automating the disease detection in chest X-rays using a new depth wise Multi-kernel Convolution (MD-Conv). Fine-grained characteristics of medical images, tissue texture and structure, are valued to be important in high-resolution medical images, and to this end this study addresses problems of varying scales of the disease, and computational complexity. As the MD-Conv module incorporates depth wise convolution kernels of different sizes in the same layer, it can effectively use different scales of features with an efficient computation cost. Their approach was tested on the Chest X-ray 14 and pediatric pneumonia datasets and reported high AUC and further improved the performance and low computational cost relative to standard CNNs.

## 3. Problem Statement

Lung cancer is a significant threat to the community since it is one of the leading causes of death in many countries due to cancer. Entirely fewer and more ineffective treatment options remain when a person is diagnosed late. Although they may be right, standard diagnostic tools such as imaging and biopsy are invasive, expensive, and in many cases, not available in early diagnosis. There exists an acute necessity to seek alternative solutions that can help in early identification and intervention. The current developments in ML revealed their potentially high potential application in medical diagnosis due to the ability to conduct data-based analytic operations on complicated clinical and behavioral points. Nevertheless, the successful use of the specified methods to predict lung cancer presupposes thorough attention to data preprocessing, feature encoding, and model selection. Moreover, the problem is to find the pattern that is connected to the possible existence of lung cancer by means of survey-based data, which can include symptoms, lifestyle preferences and demographic data. The project will assist in overcoming the obstacles by creating and assessing a survey-based ML system to identify lung cancer early. It aims at identifying suitable categorization algorithms which can reliably predict the existence of lung cancer. This will lead to the creation of non-invasive diagnostics that could provide medical practitioners with clinical decision-making and risk in time.

## 4. Proposed Method

According to survey data, the proposed solution develops a predictive model of early lung cancer diagnosis with the application of ML methods. The key steps of the methodology include data preparation, encoding and scaling, model creation and training and evaluation, and evaluation of

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0456
e ISSN: 2584-2854
Volume: 03
Issue: 09 September 2025
Page No: 2891 - 2896

performance. Categorical attributes can transform target labels into a numerical representation that will be machine readable. This is followed by standardization that ensures the removal of the biases, which are related to scales as well as ensuring that each feature plays an equal role in model learning. The five ML algorithms are used to make comparisons as DT, RF, AdaBoost, Gradient Boosting, and LR. These models are selected based on the effectiveness of performing tasks around binary classification. Training data is used to train each model, and the final evaluation is carried out by using test set where each model is scored. This comparison of individual classifier predictive powers allows one to find the most effective model to predict cases of lung cancer and later extended to identifying which type of lung cancer is possible based on patterns of symptoms using AdaBoost classifier, which once again exhibited the best performance. This improvement will allow narrower classification of lung cancer type not only by early detection but also by increasing the clinical relevance and utility of the proposed system.

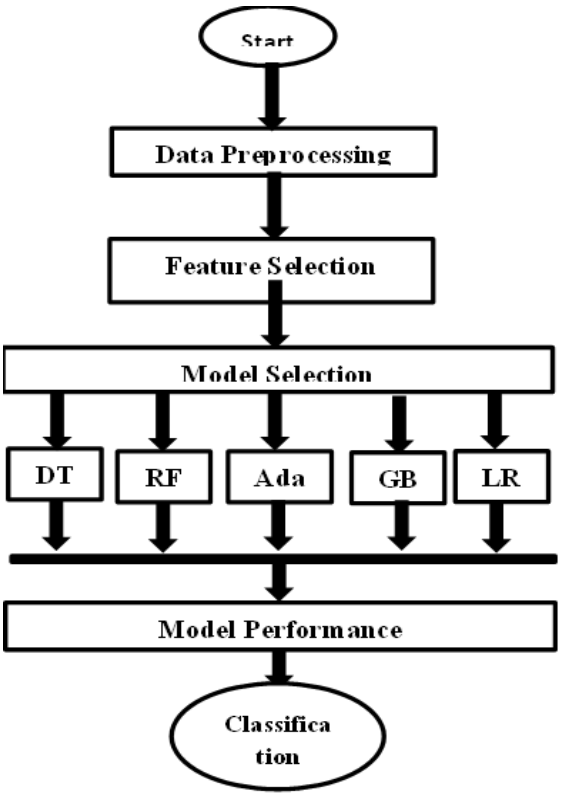


**Figure 1 Proposed Architecture**

## 5. Results And Discussion

Sixteen variables and 309 cases of lung cancer that are publicly available were used in this work to apply several supervised ML models to make lung cancer predictions. Some of the numerical and categorical variables in the data are the features such as gender, age, smoking habits, exhaustion, chest discomfort and dyspnea among others. The five trained models were used to classify data and included DT, RF, AdaBoost, Gradient Boosting, and LR. The most prevalent classification criterion, such as precision, recall, F1-score, and confusion matrix analysis, were applicable in the test of performance. The RF and Gradient Boosting algorithms are part of the ensemble-based algorithms that can reduce overfitting and generalizing data less on unknown data that is why this model exhibited higher classification power compared to the other models. The visualization of classes distribution showed an acceptable balanced set. The best-performing models were demonstrated with positive and negative confusion matrices showing a high true positive rate and the true negative rate, which indicated that the models could detect the presence of cases of lung cancer and exclude it. This approach demonstrates that ensemble models can exhibit good predictive capabilities regarding the early identification of lung cancer, particularly Random Forest and Gradient Boosting.

### 5.1. Feature Scaling – Standardization

The features were standardized using StandardAero, which transforms the dataset by subtracting the meaning and scaling it to have unit variance, ensuring all features contribute equally to the model

$$z = \frac{x - \mu}{\sigma}$$

**Where:**

- Besides exercise, Z is also commonly used to represent the standardized value or z-score in statistics that an individual data value falls above or below the mean by the number of standard deviations.
- X is the individual data point
- The feature is obtained with the use of division, that is, dividing the sum of all the observations by the total number of the

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0456
e ISSN: 2584-2854
Volume: 03
Issue: 09 September 2025
Page No: 2891 - 2896

observations produces the μ which is the average of the feature.

- σ represents the standard deviation of the feature, indicating how much the values of that feature deviate from the meaning.

### 5.2. Accuracy

The proportion of accurately anticipated cases to all predictions is known as accuracy.

$$\textbf{Accuracy} = \frac{\textbf{TP} + \textbf{TN}}{\textbf{TP} + \textbf{TN} + \textbf{FP} + \textbf{FN}}$$

**Where:**

- TP- True Positive True positives are those cases where the model is correct in predicting a positive value that is, the correct label is positive, and the model accurately predicts it as such.
- TN -True Negative is where the model made a correct prediction about the negative outcome and this is, the actual label and the model also made a positive prediction as negative.
- FP -False Positive is the result whereby, when the model takes the negative outcome as the positive one.
- FN -False Negative is a situation when the model has made a wrong prediction, i.e., the model predicts that the actual outcome is negative, when it is positive.

### 5.3. Precision

The percentage of the correct predictions in a positive direction is based on the overall percentage of the positive estimation of the predictions that constitute the precision measure. It reflects the number of positive cases out of the number of anticipated positive cases.

$$\textbf{Precision=TP/(TP+FP)}$$

### 5.4. Recall

Recall is defined as the ratio between all successful bright predictions, on the one hand, and the actual positive prediction, on the other. It points to how well models can identify TP.

$$\textbf{Recall} = \frac{\textbf{TP}}{\textbf{TP} + \textbf{FN}}$$

### 5.5. F1-Score

Even more so with unbalanced datasets, the F1-score, calculated via harmonic means of Precision and Recall, gives a more reliable appertaining to algorithm performance than accuracy.

$$\textbf{F1} - \textbf{Score} = 2 = \frac{\textbf{Precision x Recall}}{\textbf{Precision} + \textbf{Recall}}$$

### 5.6. Model Performance

Key performance indicators—were utilized to evaluate the effectiveness of five classification models: DT, RF, AdaBoost, Gradient Boosting, and LR.

**Table 1 Model Performance**

| Model | Accuracy (%) | Precision | Recall | F1-Score |
|---|---|---|---|---|
| **DT** | 93.55 | 0.98 | 0.95 | 0.96 |
| **RF** | 96.77 | 0.98 | 0.98 | 0.98 |
| **Adaboost** | 98.39 | 0.98 | 1.00 | 0.99 |
| **Gradient Boosting** | 95.16 | 0.98 | 0.96 | 0.97 |
| **LR** | 96.77 | 0.98 | 0.98 | 0.98 |

As Table 1 indicates, five ML models such as DT, RF, AdaBoost, Gradient Boosting and LR were used to predict lung cancer. The most promising outcomes were stated in AdaBoost where the accuracy equaled 98.39%, precision, 0.98, 100 percent perfect recall, and the F1-Score relied on 0.99. These results highlight the immense capacity of the AdaBoost in providing the correct answer, i.e. determining the positive and negative instances. RF and LR did not do any worse than each other, with both attaining nearly a perfect accuracy of 96.77 and F1-Score of 0.98.The Gradient Boosting ranked third with an accuracy of 95.16 and had one of the most similar competencies of all, with a F1-Score of 0.97.The DT had the highest precision at 0.98 and the lowest recall at 0.95 and the F1-Score was 0.96 and an overall accuracy record of 93.5. In figure 2, one can observe the accuracy comparison of five ML models based on their performance in classification: DT, RF, AdaBoost, Gradient Boosting, and LR. Based on the

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0456
e ISSN: 2584-2854
Volume: 03
Issue: 09 September 2025
Page No: 2891 - 2896

graph, it can be observed that AdaBoost classifier performs better in terms of accuracy. This graphical illustration supports the quantitative conclusions of Table 1 by further illustrating that AdaBoost is the most suitable model in quality of classification of the dataset of the study.

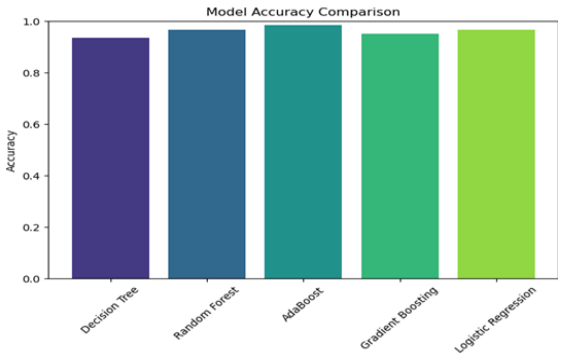


**Figure 2 Model Accuracy Comparison**

## Conclusion

This paper sheds light on the promising applications of ML in predicting lung cancer given structured survey information. When different supervised learning techniques 1 insight were applied, it was demonstrated that classification of high or low risk based on non-invasive factors like lifestyle, symptoms, demographic factors was feasible through a data driven approach. Ensemble-based models, particularly AdaBoost, were among the models to give the best results and performed very well in important metrics like accuracy (98.39%), recall (1.00), and F1-score (0.99). The suggested framework highlights the feasibility and effectiveness of ML incorporation in medical processes, particularly, in cases concerning a disease where an early diagnosing (such as lung cancer) would be paramount in survivor outlook improvement. Besides its early diagnosing, AdaBoost model was also successfully extended to classify kinds of cluster in lung cancer with the pattern of clinical and symptoms, and the accuracy of diagnosis and scheduling treatment was improved. Summing up, this paper confirms that machine learning can be a revolutionizing factor, and it should be introduced to support usual diagnostic approaches carefully, by providing the necessary preprocessing and setting up the right algorithm. The future investigation of such work can be implemented by adding various more data to it, including imaging data, to implement the models within the real-world medical settings to assess the extent to which the models can be used effectively and on scale.

## References

[1]. Alzahrani, A. (2025). Early Detection of Lung Cancer Using Predictive Modeling Incorporating CTGAN Features and Tree-Based Learning. IEEE Access.
[2]. Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD explorations newsletter, 6(1), 20-29.
[3]. Bhuiyan, M. S., Chowdhury, I. K., Haider, M., M., Jisan, A. H., Jewel, R., 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.
[4]. Wayahdi, M. R., & Ruziq, F. (2022). KNN and XGBoost Algorithms for Lung Cancer Prediction. Journal of Science Technology (JoSTec), 4(1).
[5]. Ojha, T. R. (2023). Machine learning based classification and detection of lung cancer. Journal of Artificial Intelligence and Capsule Networks, 5(2), 110-128.
[6]. Song, Y. Y., & Lu, Y. (2015). Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry.
[7]. Nasrullah, N., Sang, J., Alam, M. S., Mateen, M., Cai, B., & Hu, H. (2019). Automated lung nodule detection and classification using deep learning combined with multiple strategies. Sensors, 19(17), 3722.
[8]. Wu, Y., & Lin, L. (2020, September). Automatic lung segmentation in CT images using dilated convolution based based on weighted fully convolutional network. In Journal of Physics: Conference Series (Vol. 1646, No. 1, p. 012032). IOP Publishing.

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0456
e ISSN: 2584-2854
Volume: 03
Issue: 09 September 2025
Page No: 2891 - 2896

[9]. Irtaza, M., Ali, A., Gulzar, M., & Wali, A. (2024). Multi-label classification of lung diseases using deep learning. IEEE Access.
[10]. Hu, M., Lin, H., Fan, Z., Gao, W., Yang, L., Liu, C., & Song, Q. (2020). Learning to recognize chest-Xray images faster and more efficiently based on multi-kernel depth wise convolution. IEEE Access, 8, 37265-37274.