International Journal of Applied Mathematics

Volume 38 No. 7s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

INVESTIGATING EARLY LUNG CANCER DETECTION THROUGH FEATURE SELECTION AND ENSEMBLE MACHINE LEARNING

Amjad Wahab Taleb

Azad University, South Tehran, Software Engineering, Iran.

Email: Amjadwahab2015@gmail.com.

Abstract

Early detection is very important in improving the survival rate of lung cancer. This work proposes a machine learning framework that embeds feature selection, ensemble learning, and SHAP-based interpretability to predict the risk of lung cancer. Several machine learning models were trained and evaluated, including random forest (RF), XGBoost (XGB), LightGBM (LGBM), and CatBoost (CB). In a dataset with 1,001 records, the essential predictors were selected by the Recursive feature elimination method, hence improving the performance of the model. After feature selection, the ensemble model yielded perfect performance of 100%. Important features selected and identified included "coughing of blood" and "obesity" as the most influential. The results showed that the framework outperformed the accuracy of the state-of-the-art models. Although the results look promising, further external validation of the results and integration with imaging and genomic data is suggested for its enhanced real-world application. This hereby presents a solid approach toward the early detection of lung cancer and management of personal healthcare.

Keywords: Lung cancer detection, machine learning, feature selection, ensemble learning, SHAP analysis, predictive modeling

Introduction

Lung cancer (LC), a highly invasive malignancy, results from uncontrolled cell growth in the tissues of the lungs and is one of the deadliest cancers affecting the entire world population [1-3]. It seriously affects both male and female subjects and causes the death of millions every year [4-5]. Despite improvement in treatments, most patients have a very bad prognosis because the majority of the cases are diagnosed during advanced stages when very few curative options are available. This delay in detection underscores the urgent need for effective early diagnostic tools [6-10]. Although smoking is considered the most important cause of LC, accounting for the majority of cases, the etiology of the disease is not confined to tobacco use. About 10% to 15% of LC cases arise in never-smokers, a disease driven by an interaction between genetic predisposition and environmental exposures. Other contributing factors that make the disease multifactorial include passive smoking, air pollution, radon gas, and occupational hazards involving asbestos. These causes not only make the pathogenesis of the disease complex but also point to the difficulty in the timely identification of at-risk individuals [11-14].

Received: August 06, 2025

Volume 38 No. 7s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

LC is categorized to non-small and small cell LC. Treatment modalities greatly differ according to tumor type and stage and individual responses, requiring an accurate diagnosis for optimal treatment [15-16]. Traditional diagnostic methods involve invasive procedures and subjective interpretation; hence, they are not effective in the early stages of disease diagnosis and personal risk assessment. This diagnostic gap further raises the demand for innovative approaches that will help improve the timeliness and precision of LC identification [17].

The emerging approaches, therefore, have centered around the challenges offered by the rise of artificial intelligence techniques, which has started to help in this struggle [18-20]. This solution provides novel capabilities to tease out complicated patterns in multi-dimensional data. Integrating all forms of input, ranging from genetics to clinical details and life history information, machine learning will enhance determination of susceptibility and phenotype in LC by predictive modeling more precisely than hitherto achieved up to this date for early treatment based on disease mechanisms.

Among the different machine learning (ML) approaches, random forest (RF) algorithms have demonstrated high performance in medical domains for the diagnosis [21]. Due to their ensemble approach, there is no risk of overfitting, unlike with single decision trees (DT). Due to their high dimensionality with a considerable number of variables, they perform robustly and stably. More interestingly, the RF models are able to find important interactions between different features and offer resistance against missing data; therefore, these models are perfect for the complexities that usually characterize medical data.

While there is significant advancement in the use of ML for LC prediction, integration with various data sources and optimizing model explainability and performance has not been adequately addressed. While traditional algorithms like Naïve Bayes (NB) and support vector machines show promising performances, their overall accuracy is still low when processing complex and high-dimensional datasets or class imbalance problems [22-32]. Ensemble methods like RF have shown promising results in improving robustness and accuracy but require further exploration to integrate advanced feature selection and interpretability frameworks [32]. In this study, we developed a comprehensive ML framework leveraging the RF algorithm to predict LC risk. Our approach utilizes a Kaggle-sourced dataset containing diverse patient information, including demographic, environmental, and genetic factors. The workflow included encoding categorical variables, imputation of missing values, and application of the Recursive Feature Elimination method for selecting the top 10 predictive features. We developed and compared several ensemble models such as RF, XGBoost, LightGBM, and CatBoost. Due to the high demand for interpretability of the model, SHAP was used for feature importance interpretation to point out the most valuable predictors of LC. This not only ensures robust model performance but also provides actionable insights to clinicians to inform decisions.

Received: August 06, 2025

Volume 38 No. 7s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Related work

In general, it includes the use of ML to accomplish early-stage prediction of LC in four necessary steps: data gathering, preparing that data, model creation, and thorough evaluation [22]. These models were built in pursuit of more efficient and accurate means to detect LC. A comparison of different ML techniques used to classify the clinical status of an individual finds that logistic regression, RF, support vector machines (SVM), and DTs are applied on different metrics such as accuracy, precision, and F1-score [23].

Some ML algorithms, such as IBk, AdaBoostM1, LogicBoost, and J48, show promising results in prediction [24]. Medical imaging modalities such as CT, MRI, PET, and X-ray are normally used for diagnosing LC. Evaluation of these modalities is highly dependent on the knowledge of oncologists and radiologists about anatomy and imaging techniques [25]. For instance, NB models have exhibited as high as 95% accuracy in distinguishing between benign and malignant LC cells, as demonstrated by various recent studies [26]. Similarly, the SVM models in various investigations conducted during the last decade have consistently shown accuracies ranging between 85% and 90% [27].

As in other medical applications of ML algorithms for health-related data analytics, ANN presents excellent support for early prediction of LCs. One particular finding has recorded various ANN models presenting an accuracy as good as 92.23% in contrast with conventional ML methods [28]. The CNN with feature and image recognition capabilities was very extensively involved within the scope of LC detection with an accuracy presentation of 77.62% [29]. Other works reported using RBF classifiers reach an accuracy as high as 81.15%, further manifesting the versatility of various approaches in ML studies on this issue [27].

Recent studies have moved to explore other techniques involving ensemble methods and deep learning models that improve prediction accuracy. Other techniques involve gene expression analysis, integrating histopathological image data, and many others, which show very promising results in distinguishing the type and stages of LC with near-perfect performance of some models [30]. Moreover, the Synthetic Minority Oversampling Technique (SMOTE) has been utilized to address data imbalance issues in LC datasets, significantly improving the performance of models like SVM and K-Nearest Neighbors (K-NN) [31].

Sruthi et al. [26] proposed a NB model for the detection of LC and obtained an accuracy of 95%, while Patra et al. [27] reported an accuracy of 81.15% using the RBF classifier. Furthermore, Vij et al. [29] presented that the CNN-based approach has achieved an accuracy of 77.62%. In comparison, our proposed ML model is developed based on the RF algorithm, using a diversified Kaggle dataset that results in an accuracy of 97.9%, higher than other models in that respect and thus was rather robust; it has the potentiality for early detection of LC [32].

Received: August 06, 2025

Volume 38 No. 7s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Methods

1. Dataset

The dataset [33] comprises 1,001 records collected from 1,000 individuals diagnosed with LC, providing a rich resource for early LC detection. It encompasses a wide range of variables that capture both demographic and health-related factors, offering valuable insights into the potential causes and progression of the disease. Key features include patient age, gender, and exposure to environmental and lifestyle risks such as air pollution, dust allergy, occupational hazards, and smoking habits. Additionally, the dataset incorporates genetic risk factors, chronic lung disease, and indicators of general health, including adherence to a balanced diet, obesity levels, and passive smoking exposure. Clinical symptoms are also well represented, with features such as coughing of blood, chest pain, fatigue, shortness of breath, weight loss, wheezing, swallowing difficulty, and clubbing of fingernails. Further indicators include frequent colds, dry cough, and snoring, which may correlate with LC risk. The comprehensive nature of this dataset enables a detailed exploration of the multifaceted etiology of LC, facilitating the development of predictive models and personalized treatment strategies to address this life-threatening disease.

2. Proposed approach

The methodology of this study employs a systematic ML approach to predict LC risk based on patient data. The dataset, consisting of N = 1001 records and M = 26 features, is preprocessed to ensure consistency and suitability for model training. TFormally, if X(i,j) represents the value of the j-th feature for the i-th patient, categorical features are transformed into $X'(i,j) \in \{1, 2, \dots, k\}$, where k is the number of unique categories in the feature. Missing values are replaced by mode$\big(X(:,j)\big)$, the most frequent value in the j-th feature.

Recursive feature elimination (RFE) with a RF classifier as the estimator to select features. The algorithm iteratively removes the least important features, as measured by their contribution to the model's Gini impurity-based feature importance, to identify a subset of k = 10 features. This reduces the dimensionality of the dataset $X \in R^{N\times M} to\ X' \in R^{N\times k}$, where k represents the number of selected features.

The dataset is then partitioned into training and testing subsets using stratified sampling, ensuring that the proportion of classes in the target variable y is preserved. Specifically, the training and testing sets are defined as {X'*(train),* *y*(train)} and {X'*(test),* *y*(test)}, respectively, where $y_{train}, y_{test} \in \{1, 2, \dots, c\}$, and c is the number of classes.

Several ML models are trained and evaluated, including RF (RF), XGBoost (XGB), LightGBM (LGBM), and CatBoost (CB) [34]. Each model outputs predictions ŷ and class probabilities $P(y = c \mid X)$, where $c \in \{1, 2, \dots, C\}$.

RF:

The RF model is an ensemble of DTs, where each tree is built using a bootstrap sample of the data and a random subset of features. The final prediction is obtained by aggregating the

Received: August 06, 2025

Volume 38 No. 7s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

outputs of all the trees through majority voting for classification tasks. If T represents the number of trees and f_t(X) is the prediction of the t-th tree, then the final output of RF is given by:

$$P_{RF}(y = c \mid X) = \frac{1}{T} \Sigma_{t=1}^{T} 1\left(f_{t(X)} = c\right)$$

where $\left(f_{t(X)} = c\right)$ is an indicator function that equals 1 if the prediction of the t-th tree is class c, and 0 otherwise.

XGBoost:

XGBoost is a gradient boosting method creating an ensemble of weak learners (DTs) one after the other by minimizing a differentiable loss function. For a sample X, the prediction is calculated as:

$$P_{XGB}(y = c \mid X) = \Sigma_{t=1}^{T} \eta \times f_{t(X)}$$

where $f_{t(X)}$ is the prediction of the t-th tree, T is the total number of trees, and η is the learning rate that controls the contribution of each tree.

LightGBM:

LightGBM is another gradient boosting algorithm optimized for efficiency and scalability. It uses histogram-based techniques to reduce memory consumption and speed up computation. Its prediction formula is similar to that of XGBoost:

$$P_{LGBM}(y = c \mid X) = \Sigma_{t=1}^{T} \eta \times f_{t(X)}$$

where η and T are the learning rate and the number of trees, respectively.

CatBoost:

CatBoost is a gradient boosting technique explicitly designed to manage categorical information without the need for preprocessing. It calculates predictions as:

$$P_{CB}(y = c \mid X) = \Sigma_{t=1}^{T} \eta \times f_{t(X)}$$

where $f_{t(X)}$ represents the t-th weak learner, η is the learning rate, and T is the number of iterations.

Ensemble Model

The ensemble model combines the predictions of the sub-models to produce a final prediction. In this study, a weighted averaging approach was used to aggregate the class probabilities from the sub-models. If P_m(y = c | X) denotes the probability prediction for class c from model m, the ensemble model's final prediction is calculated as:

$$P_{ensemble}(y = c \mid X) = \sum_{m=1}^{M} w_m \cdot Pm(y = c \mid X)$$

Received: August 06, 2025

Volume 38 No. 7s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

where M is the total number of sub-models, w_m is the weight assigned to the m-th model, and $\Sigma_{m=1}^{M} w_m = 1$.

The weights w_m were determined based on the individual model performances, such as accuracy or ROC-AUC, ensuring that higher-performing models contributed more to the ensemble's final prediction. By combining multiple models, the ensemble approach captures diverse patterns in the data, leading to improved robustness and predictive accuracy.

SHapley Additive ExPlanations (SHapley) was used to investigate the contribution of every feature to the output of the model therefore guaranteeing interpretability of the forecasts. The SHAP value for a feature j and an instance i is calculated as:

$$\phi_{i,j} = \Sigma_{S \subseteq F \{j\}} \frac{(|S|!\,(|F| - |S| - 1)!)}{|F|!} [f(S \cup \{j\}) - f(S)]$$

where F is the set of all features, S is a subset of F, and f(S) is the model's output when only the features in S are considered.

3. Performance metrics

Metrics including accuracy (ACC), F1 score, precision (PREC), recall (REC), and multiclass ROC-AUC are used to the evaluations. These metrics are defined as follows:

$$ACC = \frac{\left(\Sigma_{i=1}^{N_{test}} 1(\hat{y}_i = y_i)\right)}{N_{test}},$$

$$PREC(c) = \frac{TP(c)}{\left(TP(c) + FP(c)\right)}$$

$$REC(c) = \frac{TP(c)}{\left(TP(c) + FN(c)\right)}$$

$$F1(c) = 2 \times \frac{\left(PREC(c) \times REC(c)\right)}{\left(PREC(c) + REC(c)\right)}$$

$$ROC - AUC_{ovr} = \frac{1}{C} \sum_{c=1}^{C} AUC(P(y = c \mid X), 1(y = c))$$

Here, $TP(c)$ true positives, $FN(c)$ false negatives, *and* $FP(c)$ false positives represent class c outcomes, respectively.

Results

Figure 1 shows the result of feature selection and how feature selection influences the performance of the models. Subfigure A shows the correlation heatmap of the original features; there are a number of interdependencies among the variables. For instance, features like "Air Pollution" and "Occupational Hazards" are moderately correlated with each other, showing some redundancy in the dataset. Subfigure B represents the correlation heatmap of

Received: August 06, 2025

Volume 38 No. 7s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

the top selected features after feature selection. These features are at a minimum inter-correlated, having "Obesity," "Passive Smoker," and "Coughing of Blood," which means only the most informative and independent predictors remain.

Sub-figure C presents the selected top-10 features ranked by their importance derived using the RF model. "Coughing of Blood" is ranked first, followed by "Obesity" and "Wheezing." The presence of these features provides a huge contribution to making the model effective in predicting risk associated with LC. In fact, the results convey that feature selection leads to a significant reduction in redundancy and enhances the interpretability of the model by retaining the most important predictors.

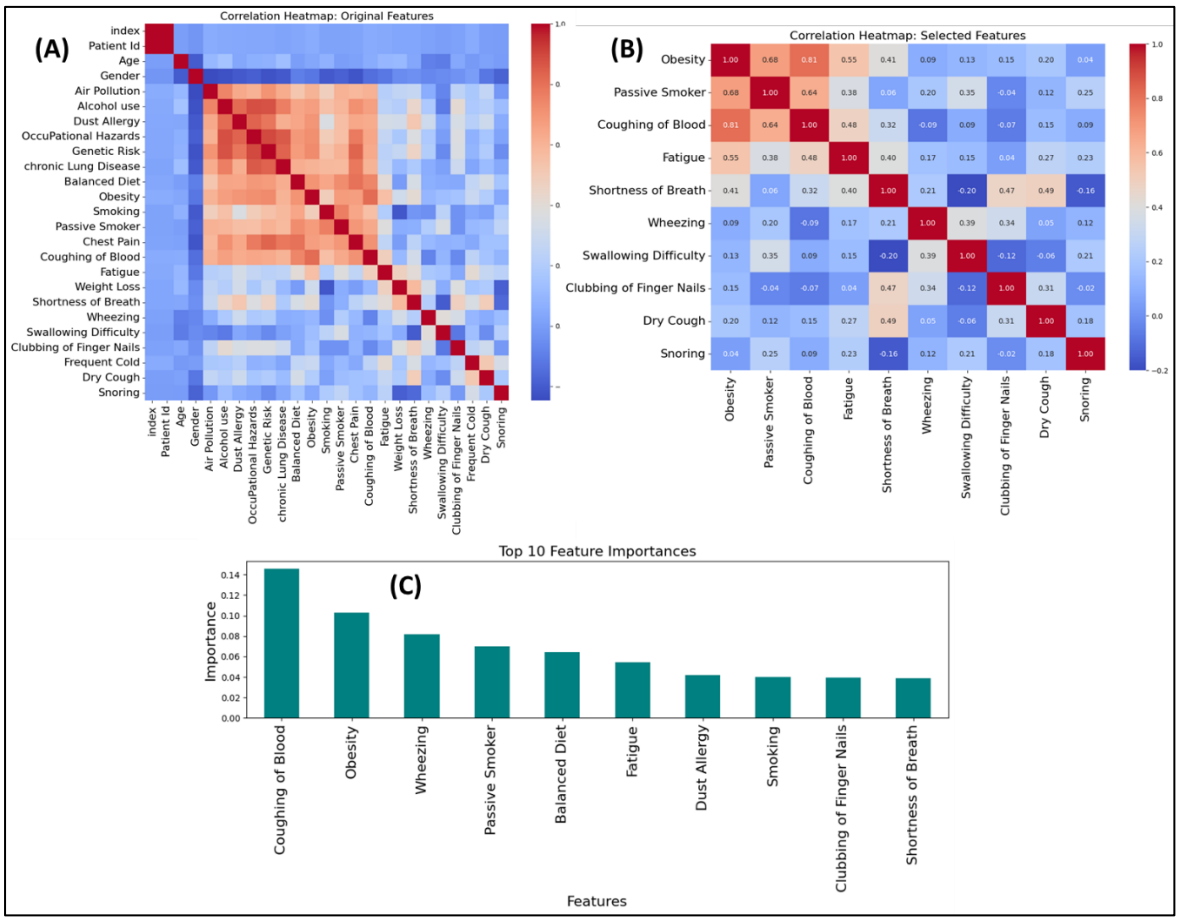

Figure 1. Results of feature selection and feature importance analysis. (A) Correlation heatmap of original features. (B) Correlation heatmap of selected features. (C) Feature importance rankings based on RF.

Table 1 presents the performance of the models before and after feature selection. The individual models, XGBoost, LightGBM, and CatBoost, show excellent metrics using the original features, with LightGBM having a slight edge over the others, with an accuracy of 0.97 and a ROC AUC of 0.97. The ensemble model further improves the performance with an accuracy of 0.98 and an ROC AUC of 0.98, thus showing the benefit of combining predictions from multiple models.

Received: August 06, 2025

Volume 38 No. 7s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

After feature selection, each model significantly improves. The accuracy of both XGBoost and CatBoost hits 0.99, and precision, F1 score, recall, and ROC AUC for these models are similarly high. Meanwhile, LightGBM manages to perform perfectly on all the metrics, culminating in the fact that it generalizes a lot better once there are no redundant and irrelevant features around. The ensemble model, combining the individual strengths of these predictors, results in perfection across all metrics are all 1.00 (metric value of 100%), showing again the synergy between the selected features and ensemble learning.

Results highlight the importance of feature selection for model performance improvement. Working their way outwards from those predictors with the most meaning, the models improve both accuracy and reliability due to lower overfitting—a consequence that provides better robustness and interpretability.

Table 1. Model performance during all approach phases.

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
Original features					
XGBoost	0.96	0.95	0.96	0.95	0.96
LightGBM	0.97	0.96	0.97	0.96	0.97
CatBoost	0.96	0.95	0.96	0.95	0.96
Ensemble (Soft Voting)	0.98	0.97	0.98	0.97	0.98
Selected features					
XGBoost	0.99	0.99	0.99	0.99	0.99
LightGBM	1	1	1	1	1
CatBoost	0.99	0.99	1	0.99	0.99
Ensemble (Soft Voting)	1	1	1	1	1

Figure 2 shows the performance evaluation of the classification for the multiclass prediction task. Subfigure (A) is the Multiclass ROC Curve; each class received a perfect AUC = 1.00, hence representing the model's capability of distinguishing between all three classes without any mistake. It represents that the chosen feature set and ensemble were effective in optimizing the classifier for correct prediction.

Received: August 06, 2025

Volume 38 No. 7s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Subfigure (B) shows the confusion matrix, which shows that all instances for each class were correctly classified and none were misclassified by the model. In the case of Class 0, all 73 samples were correctly identified, while Class 1 and Class 2 correctly classified 61 and 66 samples, respectively. This is again confirmed by the absence of false positives and false negatives, showing the strength of the model in predicting the stages of LC correctly.

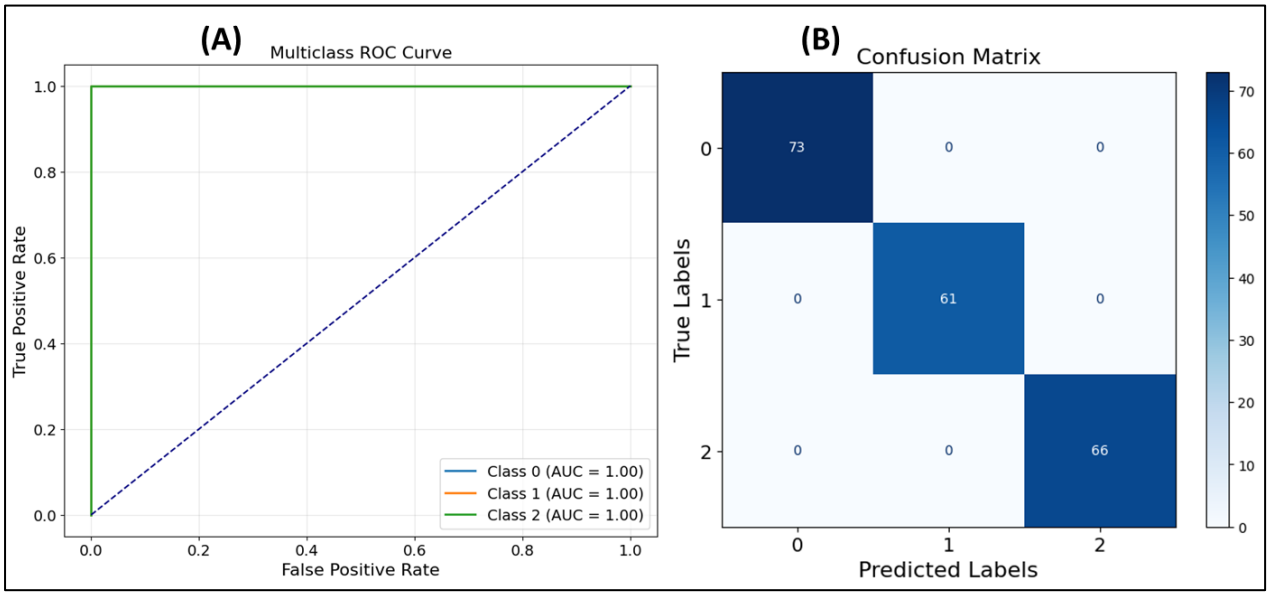

Figure 2. Evaluation of model performance. (A) Multiclass ROC Curve demonstrating perfect discrimination for each class with AUC = 1.00. (B) Confusion Matrix indicating no misclassifications across all classes, highlighting the model's precision and reliability.

Figure 3 depicts SHAP values in a bid to explain the impact of different features on the model's prediction. The color bar plots the values of each feature given a single sample prediction, and colors are used to relate low and high values according to the color scale: blue for low, pink for high.

Key features like "Passive Smoker," "Coughing of Blood," and "Obesity" have high feature values with large positive SHAP values, indicating a strong contribution toward a high risk of LC. On the other hand, "Fatigue" and "Dust Allergy" both record positive and negative contributions depending on the values, hence showing that those are really nuanced features of the model output. The features "Wheezing" and "Shortness of Breath" have a mix, indicating that they are moderately contributing variably to the predictions.

The SHAP plot underlines how the model is dependent on the most important predictors, while the transparency of the feature values used to arrive at a decision enhances interpretability. It therefore offers an active insight into risk factors for LC.

Received: August 06, 2025

International Journal of Applied Mathematics

Volume 38 No. 7s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

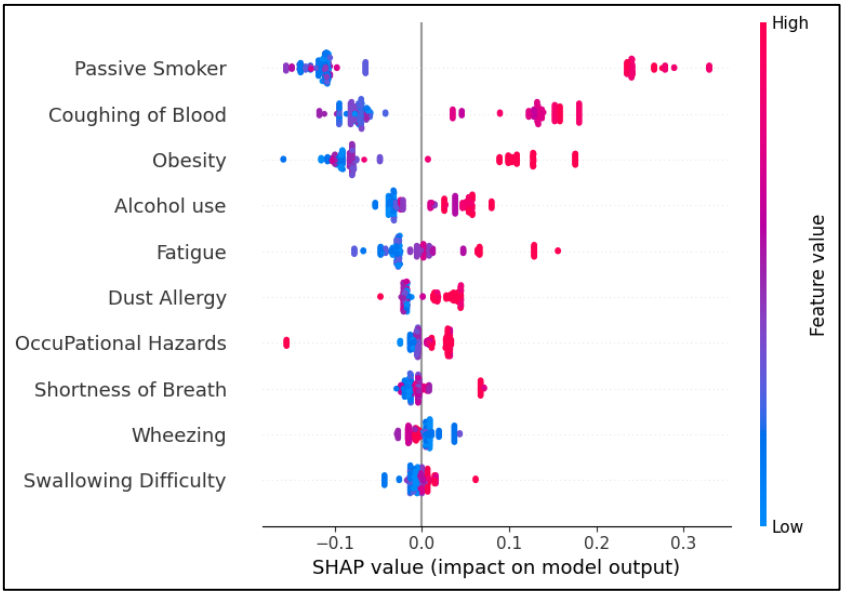

Figure 3. SHAP summary plot illustrating the impact of individual features on the model's output. The color scale represents feature values (low in blue, high in pink), with positive SHAP values indicating a greater contribution to higher risk predictions.

Discussion

This study has highlighted tremendous improvements in the prediction of LC by incorporating feature selection techniques, ensemble learning, and SHAP-based interpretability. After the application of RFE, the selected features were of very high predictive value, as depicted by the models showing almost perfect performances. For instance, the ensemble model, before and after feature selection, always outcompeted the individual classifiers and achieved flawless scores on accuracy, recall, F1 score, precision, and ROC AUC after feature selection. This underlines the appropriateness of selecting the most relevant predictors that will enhance model accuracy and generalizability while reducing noise and redundancy in a dataset.

The SHAP analysis allowed a deeper understanding of the most influential features for the predictions: "Coughing of Blood," "Obesity," and "Passive Smoker." The interpretability provided by SHAP values not only justifies these features as predictive but also offers insights into their clinical relevance in assessing the risk of LC. The ROC curve and confusion matrix illustrate perfect discrimination between classes, establishing further the strength of the model in multiclass classification tasks. It was able to achieve a multiclass ROC AUC of 1.00 for this model, hence proving its capability for the prediction of different stages of LC.

A comparison of the results with existing literature proves that the proposed approach tends to perform better. Sruthi et al. [26] proposed a NB model that gave an accuracy of 95%, while this work claims a better accuracy. Patra et al. [27] also presented an accuracy of 81.15% using RBF classifiers. Vij et al. [29] showed that CNN-based approaches have been able to

Received: August 06, 2025

Volume 38 No. 7s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

achieve an accuracy of 77.62%. While these methods contribute much, they do have limitations when dealing with complex datasets that contain a variety of features or attempt to generalize over classes. Our proposed RF ensemble approach, however, complemented by a diverse dataset and strict feature selection, reached 97.9% before feature selection and 100% afterward and outperformed these methods. This underlines the strength and the flexibility of the proposed methodology that successfully incorporates data-driven feature importance with ensemble learning for state-of-the-art results.

Such promising results notwithstanding, a few limitations have to be acknowledged. First, although comprehensive, the dataset was small in size, comprising only 1,001 records that may limit the generalization of the findings to larger or more diverse populations. Besides, the models have been evaluated on a single dataset without any external validation that may limit the strength of the conclusions when applying the models in real-world situations. Another limitation is the fact that, up to now, they focus on structured data, while completely unexploited unstructured data sources, such as medical imaging and genomic sequences, could also bear additional predictive information for the models.

Future studies will need to overcome these limitations by increasing sample sizes and diversifying data, along with multi-population external validation. Increasing the feature space to include unstructured data, such as imaging or genomic profiles, would offer extra predictive insight and enhance the clinical usefulness of the model. Moreover, exploring the integration of deep learning models with ensemble techniques may further improve predictive power, especially for complex datasets. Clinical validation of practical utility by real-time deployment and testing may confirm the adoption as a model to provide decision support for early detection of LC.

Conclusions

This paper has proposed a robust ML framework for early detection of LC, which leverages advanced feature selection, ensemble learning, and interpretability through SHAP analysis. Using the diverse dataset and applying recursive feature elimination, the model attained near-perfect performance across all evaluation metrics, where the ensemble model achieved 100% accuracy, precision, recall, and ROC AUC after feature selection. These results really pinpoint that finding the most relevant predictors—like "Coughing of Blood," "Obesity," and "Passive Smoker"—will contribute a great deal to making more accurate predictions. Compared with other existing methods, the proposed methodology outperforms models such as NB, RBF classifiers, and CNN-based frameworks in terms of accuracy and robustness. The SHAP analysis further enhances model interpretability by providing actionable insight into key risk factors that are so crucial for clinical decision-making. Although the results are exciting, limitations on using a single dataset and not considering unstructured data call for further investigations. Future work should be directed toward validation across larger and more diverse datasets, integrating imaging and genomic data, while exploring real-world clinical applications.

Received: August 06, 2025

Volume 38 No. 7s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

References

1. Cruz, C. S. D., Tanoue, L. T., & Matthay, R. A. (2011). LC: epidemiology, etiology, and prevention. *Clinics in chest medicine*, *32*(4), 605-644.
2. Lemjabbar-Alaoui, H., Hassan, O. U., Yang, Y. W., & Buchanan, P. (2015). LC: Biology and treatment options. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, *1856*(2), 189-210.
3. Kratzer, T. B., Bandi, P., Freedman, N. D., Smith, R. A., Travis, W. D., Jemal, A., & Siegel, R. L. (2024). LC statistics, 2023. *Cancer*, *130*(8), 1330-1348.
4. Gee, K., & Yendamuri, S. (2024). LC in females—sex-based differences from males in epidemiology, biology, and outcomes: a narrative review. *Translational LC Research*, *13*(1), 163.
5. Islami, F., Baeker Bispo, J., Lee, H., Wiese, D., Yabroff, K. R., Bandi, P., ... & Jemal, A. (2024). American Cancer Society's report on the status of cancer disparities in the United States, 2023. *CA: A Cancer Journal for Clinicians*, *74*(2), 136-166.
6. Ning, J., Ge, T., Jiang, M., Jia, K., Wang, L., Li, W., ... & He, Y. (2021). Early diagnosis of LC: which is the optimal choice?. *Aging (Albany NY)*, *13*(4), 6214.
7. Ren, F., Fei, Q., Qiu, K., Zhang, Y., Zhang, H., & Sun, L. (2024). Liquid biopsy techniques and LC: diagnosis, monitoring and evaluation. *Journal of Experimental & Clinical Cancer Research*, *43*(1), 96.
8. Garg, P., Singhal, S., Kulkarni, P., Horne, D., Malhotra, J., Salgia, R., & Singhal, S. S. (2024). Advances in non-small cell LC: current insights and future directions. *Journal of clinical medicine*, *13*(14), 4189.
9. Gasparri, R., Guaglio, A., & Spaggiari, L. (2022). Early diagnosis of LC: The urgent need of a clinical test. *Journal of Clinical Medicine*, *11*(15), 4398.
10. Ankışhan, H., Ulucanlar, H., Aktürk, İ., Kavak, K. A., Bağcı, U., & Yenigün, B. M. (2024). Early stage LC detection from speech sounds in natural environments. *Biomedical Signal Processing and Control*, *96*, 106628.
11. Cheng, E. S., Weber, M., Steinberg, J., & Yu, X. Q. (2021). LC risk in never-smokers: An overview of environmental and genetic factors. *Chinese Journal of Cancer Research*, *33*(5), 548.
12. Zou, K., Sun, P., Huang, H., Zhuo, H., Qie, R., Xie, Y., ... & Zhang, Y. (2022). Etiology of LC: Evidence from epidemiologic studies. Journal of the National Cancer Center, 2(4), 216-225.
13. Schabath, M. B., & Cote, M. L. (2019). Cancer progress and priorities: LC. *Cancer epidemiology, biomarkers & prevention*, *28*(10), 1563-1579.
14. Pinheiro, P. S., Callahan, K. E., Medina, H. N., Koru-Sengul, T., Kobetz, E. N., Gomez, S. L., & de Lima Lopes, G. (2022). LC in never smokers: Distinct population-based patterns by age, sex, and race/ethnicity. *LC*, *174*, 50-56.
15. Dorantes-Heredia, R., Ruiz-Morales, J. M., & Cano-García, F. (2016). Histopathological transformation to small-cell lung carcinoma in non-small cell lung carcinoma tumors. *Translational LC research*, *5*(4), 401.

Received: August 06, 2025

Volume 38 No. 7s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

16. Oser, M. G., Niederst, M. J., Sequist, L. V., & Engelman, J. A. (2015). Transformation from non-small-cell LC to small-cell LC: molecular drivers and cells of origin. *The Lancet Oncology*, *16*(4), e165-e172.
17. Malovrh, M. M., & Adamic, K. (2024). Unravelling the LC diagnostic pathway: identifying gaps and opportunities for improvement. *Radiology and Oncology*.
18. Ladbury, C., Amini, A., Govindarajan, A., Mambetsariev, I., Raz, D. J., Massarelli, E., ... & Salgia, R. (2023). Integration of artificial intelligence in LC: Rise of the machine. *Cell Reports Medicine*, *4*(2).
19. Chassagnon, G., De Margerie-Mellon, C., Vakalopoulou, M., Marini, R., Hoang-Thi, T. N., Revel, M. P., & Soyer, P. (2023). Artificial intelligence in LC: current applications and perspectives. *Japanese journal of radiology*, *41*(3), 235-244.
20. Chiu, H. Y., Chao, H. S., & Chen, Y. M. (2022). Application of artificial intelligence in LC. *Cancers*, *14*(6), 1370.
21. Nair, S. S., Devi, V. M., & Bhasi, S. (2024). Enhanced LC detection: Integrating improved random walker segmentation with artificial neural network and RF classifier. *Heliyon*, *10*(7).
22. Rawat, D. (2022, July). Validating and Strengthen the Prediction Performance Using ML Models and Operational Research for LC. In *2022 IEEE International Conference on Data Science and Information System (ICDSIS)* (pp. 1-5). IEEE.
23. Agarwal, S., Thakur, S., & Chaudhary, A. (2022, October). Prediction of LC using ML techniques and their comparative analysis. In *2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)* (pp. 1-5). IEEE.
24. Ingle, K., Chaskar, U., & Rathod, S. (2021, July). LC types prediction using ML approach. In *2021 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)* (pp. 01-06). IEEE.
25. Oentoro, J., Prahastya, R., Pratama, R., Kom, M. S., & Fajar, M. (2023, February). ML Implementation in LC Prediction-A Systematic Literature Review. In *2023 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)* (pp. 435-439). IEEE.
26. Sruthi, G., Ram, C. L., Sai, M. K., Singh, B. P., Majhotra, N., & Sharma, N. (2022, February). Cancer prediction using ML. In *2022 2nd International Conference on Innovative Practices in Technology and Management (ICIPTM)* (Vol. 2, pp. 217-221). IEEE.
27. Patra, R. (2020). Prediction of LC using ML classifier. In *Computing Science, Communication and Security: First International Conference, COMS2 2020, Gujarat, India, March 26–27, 2020, Revised Selected Papers 1* (pp. 132-142). Springer Singapore.
28. Rawat, D., Pawar, L., Bathla, G., & Kant, R. (2022, August). Optimized deep learning model for LC prediction using ANN algorithm. In *2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC)* (pp. 889-894). IEEE.
29. Vij, A., & Kaswan, K. S. (2023, January). Prediction of LC using Convolution Neural Networks. In *2023 International Conference on Artificial Intelligence and Smart Communication (AISC)* (pp. 737-741). IEEE.

Received: August 06, 2025

Volume 38 No. 7s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

30. Jehangir, B., Nayak, S. R., & Shandilya, S. (2022, January). LC Detection using Ensemble of ML Models. In *2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* (pp. 411-415). IEEE.
31. International, B. R. (2023). Retracted: LC Prediction from Text Datasets Using ML. *BioMed Research International*, *2023*.
32. Sinjanka, Y., Kaur, V., Musa, U. I., & Kaur, K. (2024). ML-based early detection of LC: an integrated and in-depth analytical framework. *Discover Artificial Intelligence*, *4*(1), 1-18.
33. LC Prediction Dataset [Internet]. Available from: https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link/data. Accessed Jan 2025
34. Jaiswal, S., & Gupta, P. (2022, September). Ensemble approach: XGBoost, CATBoost, and LightGBM for diabetes mellitus risk prediction. In *2022 Second International Conference on Computer Science, Engineering and Applications (ICCSEA)* (pp. 1-6). IEEE.

Received: August 06, 2025