Proceedings of the 2025 12th International Conference on Computing for Sustainable Global Development (INDIACom)
Bharati Vidyapeeth's Institute of Computer Applications and Management (BVICAM), New Delhi (INDIA)

# Predicting Lung Cancer using Ensemble Machine Learning Algorithms: A Comprehensive Study

Kartik Deshmukh
*Department of Computer Engineering*
*K J Somaiya Institute of Technology*
Sion, Mumbai, India
kartik.sd@somaiya.edu

Yohan Gala
*Department of Computer Engineering*
*K J Somaiya Institute of Technology*
Sion, Mumbai, India
yohan.gala@somaiya.edu

Shubham Darji
*Department of Computer Engineering*
*K J Somaiya Institute of Technology*
Sion, Mumbai, India
shubham.darji@somaiya.edu

Pradnya Patil
*Department of Computer Engineering*
*K J Somaiya Institute of Technology*
Sion, Mumbai, India
pradnya08@somaiya.edu

***Abstract*- Lung cancer is the leading cause of death from cancers. As such, it demands an earlier and more accurate diagnosis. The current study aims to unearth the different algorithm's performance that's applied for machine learning (ML) for predicting lung cancer-Logistic Regression (LR), Decision Trees (DT), K-Nearest Neighbors (KNN), Naive Bayes, Support Vector Classifier (SVC), Random Forest (RF), Gradient Boosting (GB), and Extreme Gradient Boosting (XGBoost). A clinical feature and symptoms dataset of 309 samples with 16 columns is used for lung cancer prediction. Here, a Voting Classifier has been employed as an ensemble approach wherein a few models are summed up to acquire one output. XGBoost provided a maximum accuracy of 96.64% on this dataset. In the case of the Voting Classifier, an overall accuracy of 95.80% was obtained with a cross-validation of 94.13%. It also depicts how techniques of ensemble learning are generally very efficient in enhancing the precision and reliability of predictions over lung cancer.**

***Keywords- Machine learning algorithms, Accuracy, Cross-validation, XGBoost, Ensemble learning, Lung cancer prediction***

## I. INTRODUCTION

Lung cancer is one of the most common and lethal forms of cancer worldwide, resulting in a significant proportion of global cancer-related deaths [1]. Even with treatment, the early detection of lung cancer is critical because this kind of cancer often only displays signs in its later stages. Early-stage detection can even improve the five-year survival rate from about 10 percent when diagnosed in later stages to more than 50 percent when caught early. Again, this highlights the use of diagnostic screening in cases of lung cancer, given that it may appear quite late in the condition before any symptoms are developed, such as in [2,3]. So far, conventional screening diagnostic methods such as low-dose computed tomography or LDCT have been effective in demonstrating early-stage lung cancer amongst high-risk populations. With such effective outcomes, inherent limitations regarding high false positives and risks associated with radiation are seen with LDCT. All these aspects point towards a desire for supplementary or alternative detection methods [4,5]. Machine learning (ML) methods now offer promising alternatives to cancer detection and prediction. Because of their ability to make predictions based on complex patterns in large datasets, these models can provide insights unavailable from traditional statistical methods and pinpoint subtle signs and correlations linked to early lung cancer risk [6,7]. It incorporates widely applied ML algorithms within its architecture like logistic regression, SVM, decision tree, and ensemble methods all offering unique strengths to both classification and prediction tasks. Highly advanced models such as deep neural networks and ensemble-based methods such as random forest, and XGBoost have gained tremendous performance in predictive assignments and some reports had as high as 95% accuracy in the classifications of lung cancer [11,12].

Besides, pre-processing such as Contrast Limited Adaptive Histogram Equalization (CLAHE) and segmentation methods such as Membership guided Intensity Boosted FCM using SDAM (MIBFS) clustering or bounding box-based lung extraction were found to significantly improve detection and are, therefore, suitable for high-scale application in diagnostic pipelines [26,31]. In addition, integration of clinical, imaging, and genomic information supports end-to-end predictive models with maximum diagnosis and treatment protocol accuracy and person-specific protocols customized to individuals [15,16,30].

Recent research has investigated new AI techniques for the diagnosis of lung disease. For instance, CBMIR medical image retrieval systems based on CNNs have exhibited excellent precision and recall enhancement of lung disease classification [32]. In addition, hierarchical autoencoders-based hybrid models for analysis of multi-omics data have facilitated prognosis prediction of lung adenocarcinoma based on identification of individual subgroups of patients with variable survival rates [30]. Methods such as KNN classification and K-means clustering have also been utilized for improved cancer detection diagnostic accuracy [27,28]. Moreover, image processing techniques such as HE, MMCS, and CLAHE have enhanced accuracy of deep neural network-based lung disease classification, validating preprocessing necessity for optimal model performance [29].

Considering these opportunities and challenges, this research uses a well-balanced dataset with features and symptoms to predict lung cancer risk. Based on a combination of ML algorithms from traditional and modern approaches, we can find improved accuracy and robustness of lung cancer detection. It will help increase machine learning algorithms in medical diagnostics [17,18].

This research paper is organized as follows: Section II: Literature Survey reviews existing research on machine learning techniques for lung cancer detection, emphasizing advancements, challenges, and key findings. Section III: Methodology describes the proposed approach, including data preprocessing, feature selection, and implementing various machine learning models. Section IV: Results The results of the experiments are presented with the performance analysis of the models with accuracy, precision, and recall. Finally, the Conclusion will be covering the significance of the study findings, the limitations of this research, and possible further studies in improving lung cancer detection and prediction.

## II. Literature Survey

In the last few years, a lot of research has been conducted on the use of ML models (as depicted in Table I) for the detection and predicting lungs cancer for enhancing the accuracy of diagnosis. Lung cancer is characterized by a high mortality rate, mainly because it is diagnosed late, so there is a critical need for early detection techniques [1]. Conventional diagnostics like low-dose computed tomography were very effective in the detection of lung cancer in high-risk populations but had their disadvantages such as a high false-positive rate and exposure to risks from radiation [2,3].

### *A. Machine Learning in Lung Cancer Detection*

The prediction of lung cancer has been increasingly used ML models due to the fact that it can handle large volumes of data and detect subtle, nonlinear data patterns. The clinical, demographic, and lifestyle-based lung cancer prediction has also been done using logistic regression, decision trees, SVMs, and neural networks [4, 5]. For instance, SVMs have also been applied in study experiments that obtained excellent sensitivities and specificities to identify the cancer by lung cancer clinical dataset [6].

There is other decision tree-based algorithm methods and ensemble approach methodologies such as random forest also, that can classify patient disease status with lung cancer as lung cancer patient and which further interpretable with good robust performance for that method [7,8]. The most advanced ML techniques, including deep learning, have increased the accuracy of predictions. For instance, CNNs have been used to classify lung nodules from imaging data, with an accuracy of up to 95% in classifying malignant versus benign nodules [9]. According to studies, ensemble approaches that include gradient boosting approaches such as XGBoost combined with random forest models come together to produce combinations of predictions from various algorithms. This approach improves upon the predictive accuracy, and in some cases, the results have been reported to be as good as 97% [10,11]. Such models would perform well not only concerning classification tasks but also in survival prediction, where an understanding of the risk assessment would also help in the earliest intervention for patients who face critical levels of risk [12].

### *B. Data Imbalance and Preprocessing Techniques*

One of the biggest challenges in lung cancer prediction is the problem of class imbalance where the number of cases is much less than the non-cancer cases. This tends to lead to biased predictions and poor model performance. During data preprocessing, there is encoding of categorical variables and handling of missing data for normalization of numerical features. This enables optimal performance from ML models [15].

### *C. Integrating Multi-Omics and Clinical Data*

Some studies have begun integrating multi-omics data such as genomic, proteomic, and transcriptomic data—with clinical data to provide a more comprehensive basis for prediction models. By incorporating such diverse data sources, ML models can capture complex interactions within biological systems, enhancing their predictive accuracy for lung cancer [16,17]. These multi-omics approaches have shown potential in improving early detection and tailoring personalized treatment plans, although they present additional computational challenges and require larger datasets for effective training [18].

### *D. Evaluation Metrics and Model Interpretability*

The performance of the ML model in lung cancer prediction can be measured using metrics such as accuracy, precision, recall, and F1 score, giving insight into the model performance in different dimensions [19]. Learning curves and confusion matrices are also employed to provide visual depictions of the model's performance and suggest areas for improvement. However, despite a very high accuracy, one important aspect has been model interpretability, especially in medical applications, as the understanding of contributions made by features in a prediction would be very critical. Recently, techniques like Shapley additive explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) have been advanced to improve the interpretability of complex ML models for practitioners to understand and then trust model outputs in a clinical setting [20,21,22].

### *E. Challenges*

Although the models hold promise, applications in clinical practice are still some ways off due to several key challenges, including data privacy and generalizability of the models [23]. Integrating imaging with clinical data also complicates data fusion and increases the computational complexity [24]. The need to overcome all these is therefore critical if the reliability of machine learning for predicting lung cancer needs to improve. Recent developments in image processing, clustering, and object detection have contributed to the field. Techniques such as lung segmentation through clustering and advanced neural networks [25], and deep learning models for the detection of abnormal lungs from X-ray images [26], continue to improve the accuracy of diagnostics. Other research has shown the application of K-NN classifiers, the use of clustering methods [27], and some modifications to those algorithms for better prediction accuracy [28].

Image enhancement for lung disease diagnosis with pre-trained deep neural networks by transfer learning approach

 ISSN 0973-7529; ISBN 978-93-80544-58-8 

shows very promising results [29]. Integrating multi-omics data into machine learning models would allow better subtyping and survival estimation in lung adenocarcinoma [30]. Novel clustering and neural network-based methods of lung cancer segmentation manage to achieve high accuracy with a low rate of misclassification [31].

TABLE I. SUMMARY OF RESEARCH PAPERS ON MACHINE LEARNING FOR LUNG CANCER DETECTION

| Reference | Focus Area | ML Technique | Dataset | Key Finding |
|---|---|---|---|---|
| [4,5] | Predictive modeling | Logistic Regression, Decision Trees, SVM | Clinical and demographic data | High sensitivity and specificity using SVMs |
| [6] | Prediction of lung cancer | SVM | Clinicazl Data | Improved accuracy for early-stage lung cancer |
| [7,8] | Classificatiion of Lung Cancer | Random Forest, Decision Trees | Imaging and clinical data | Robust classification and interpretability |
| [9] | Nodule classification | CNN | CT imaging data | Achieved up to 95% accuracy |
| [10,11] | Ensemble modelling | XGBoost, Random Forest | Multi-source datasets | Ensemble models achieved up to 97% accuracy |
| [12] | Risk assessment and survival prediction | Gradient Boosting | Patient survival data | Improved risk stratification for critical patients |
| [13,14] | Handling data imbalance | SMOTE | Synthetic and imbalanced datasets | Balanced dataset improves generalization |
| [16,17] | Multi-omics integration | Hybrid ML Techniques | Genomic, proteomic, and clinical data | Enhanced predictive accuracy through diverse datasets |
| [20,21,22] | Model interpretability | SHAP, LIME | ML models for lung cancer prediction | Increased trust and transparency in clinical settings |

## III. METHODOLOGY

This paper makes use of a pipeline of machine learning for preprocessing of data, training of the model, performance evaluation on several lung cancer prediction models. Data pre-processing is accompanied by feature engineering, class balancing, model selection, hyperparameter tuning, and finally performance evaluation. Ten algorithms for machine learning have been implemented and integrated into the use of an ensemble voting classifier to vote the different predictions for the output of interest.

### *A. Dataset Description*

The dataset for this job contains 309 observations and 16 variables: clinical characteristics and symptoms of lung cancer [7]. The response variable, 'LUNG_CANCER,' is binary representing the presence ('YES') or absence of lung cancer.

There are numerical as well as categorical variables kinds: gender, age, and smoking habits, amongst others- coughing, wheezing, and chest pain. Table II contains the summary, description, and the type of data. These features are critical in determining the potential predictive power of the models employed.

### *B. Data Preprocessing*

In Figure 1, data preprocessing includes multiple steps to prepare the dataset for accurate lung cancer prediction: The preliminary analysis of the dataset revealed the presence of duplicate records and missing values

Duplicate entries were removed to prevent any potential bias in the dataset, while no imputation was required as there were no missing values.

The dataset contains binary categorical variables, such as GENDER, SMOKING, and YELLOW_FINGERS, which were converted into numerical representations (e.g., 0 for "NO" and 1 for "YES") using a Label Encoder.

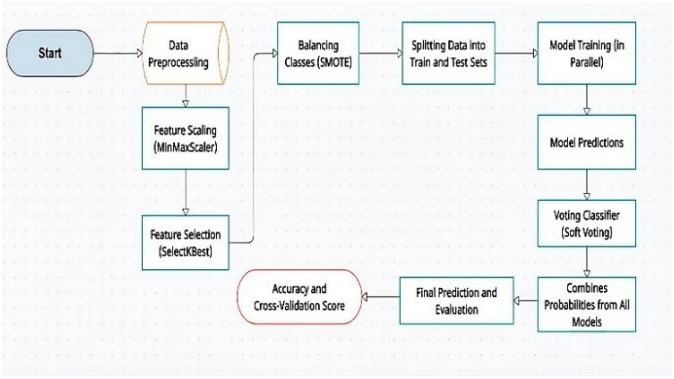


Fig. 1. Flowchart Handling Duplicates and Missing Values

This encoding facilitates the processing of categorical variables by machine learning algorithms, which require numerical inputs for optimal performance.

Standardize the feature scales, MinMaxScaler was applied to scale all variables within a [0, 1] range. Feature scaling is required by the algorithms sensitive to the magnitude of the features, like SVM and KNN. Another important aspect of feature engineering is enrichment; it created an interaction term: ANXYELFIN as the product of ANXIETY and YELLOW_FINGERS.

It aims at trying to capture any potential interaction which could enhance the power of the model in terms of predictions. Since the dataset was imbalanced concerning lung cancer-positive cases (Table II), it used ADASYN to generate synthetic samples for a minority class. ADASYN places more emphasis on harder-to-classify samples; thus, it balances the dataset and aids in recall for predictions of a lung cancer-positive case, thereby making sure that the model can learn both classes equally well, especially improving its sensitivity about lung cancer cases.

TABLE II. LUNG CANCER RISK FACTORS

| Feature | Description | Type |
|---|---|---|
| GENDER | Patient's gender (Male/Female) | Categorical |
| AGE | Patient's Age in years | Numerical |
| SMOKING | Indicates the status of smoking (0: Non-smoker, 1: smoker) | Numerical |
| YELLOW FINGERS | If yellow fingers are there or not (0: No, 1: Yes) | Numerical |
| ANXIETY | If anxiety is there or not (0: No, 1: Yes) | Numerical |
| PEER PRESSURES | If peer pressure is there or not (0: No, 1: Yes) | Numerical |
| CHRONIC DISEASES | If chronic disease is there or not (0: No, 1: Yes) | Numerical |
| FATIGUE | If experiencing fatigue (0: No, 1: Yes) | Numerical |
| ALLERGY | Are there any allergies (0: No, 1: Yes) | Numerical |
| WHEEZING | Is their wheezing (0: No, 1: Yes) | Numerical |
| ALCOHOL CONSUMING | Do you consume alcohol (0: No, 1: Yes) | Numerical |
| COUGHING | Is there coughing (0: No, 1: Yes) | Numerical |
| SHORTNESS OF BREATH | Do you have shortness of breath (0: No, 1: Yes) | Numerical |
| SWALLOWING DIFFICULTY | Is swallowing difficuly (0: No, 1: Yes) | Numerical |
| CHEST PAINS | Any chest pain (0: No, 1: Yes) | Numerical |
| LUNG CANCER | Target variable if lung cancer present or not (0: No, 1: Yes) | Categorical |

### C. Model Selection and Hyperparameter Tuning

Ten various machine learning models apply diversified algorithmic approaches and after that hyperparameter tuning for configurations by using Randomized SearchC V. The various models applied are as mentioned below:

*1) Logistic Regression:*

The simplicity of Logistic Regression and its interpretability qualify it as a baseline model. Hyperparameters, primarily the regularization parameters are tuned to make good performance with good control over overfitting.

*2) Decision Tree Classifier:*

Decision Trees are chosen as they provide interpretability and can handle nonlinear relationships. Parameters such as tree depth, minimum samples per leaf were set to optimize it so as not to overfit and generalize instead.

*3) K-Nearest Neighbors (KNN):*

KNN, the distance-based algorithm, classify cases based on the most relevant neighbors. The important parameters include the number of neighbors as well as the distance metric to be used, both tuned for optimal performance.

*4) Gaussian Naive Bayes:*

The Gaussian Naive Bayes assumes feature independence and is appropriate for small datasets. Due to its computational efficiency and robustness with continuous features, this probabilistic classifier as applied

*5) Multinomial Naive Bayes:*

Multinomial Naive Bayes is tailored for categorical features and count data. It was included in the alternative to Gaussian Naive Bayes to check the difference in performance when features are transformed.

*6) Support Vector Classifier (SVC):*

SVC with an RBF kernel was selected. It's effective in very high dimensional feature spaces. A few of the hyperparameters-tuning included regularization (C) and kernel coefficient-gamma to get the most optimal boundaries for classification.

*7) Random Forest Classifier:*

SVC with radial basis function (RBF) kernel as it works best in such high dimensional feature spaces were used. Hyperparameters set include regularization (C), and gamma of the Kernel.

*8) Gradient Boosting Classifier:*

. Gradient Boosting performs iterative improvement of weak learners. Each tree tries to correct the mistakes of the previous one. Learning rate and maximum depth were parameters set to balance bias and variance

*9) Extreme Gradient Boosting (XGBoost):*

XGBoost was an implementation of the gradient-boosting algorithm to leverage accuracy as well as be effective when the dataset became imbalanced. The principal parameters including the learning rate, max depth, and number of estimators are optimized towards high performance.

*10) Multi-layer Perceptron (MLP) Classifier:*

The MLP neural network is selected because it can learn and represent patterns which is complex in data. The accuracy of the network can be improved to a great extent by fine-tuning parameters such as the number of hidden layers, units in each layer, and activation functions. It is strong in modeling non-linear relationships between variables, hence showing better performance.

### D. Ensemble Voting Classifier

To calibrate the accuracy of the predictions generated, an ensemble voting classifier was used. For the prediction to be generated at a superior level of performance, ten independent models were integrated together. In using the ensemble, the soft voting method adopted the average value of the predicted probability coming from each model.

The soft voting technique is of greater advantage whenever individual models used vary with respect to how confident they are in the outputs produced. It weighs the predicted values by probability and hence creates a well-balanced output from the different models. This ensemble voting classifier has the strengths of each model and thus is more stable, often outperforming the single model.

### E. Model Evaluation Metrics

Each of the models, including the ensemble voting classifier, was evaluated using several key metrics to ensure a comprehensive assessment. Accuracy indicates the overall correctness of predictions across both lung cancer-positive and negative classes, providing a high-level view of model performance. However, accuracy alone can sometimes be misleading, especially in imbalanced datasets, as it may favor the majority class disproportionately.

To this effect, Precision was computed to determine how well the model could reduce false positives, with the number of true positives divided by all predicted positives. High precision is particularly critical for clinical applications since it will minimize the chance of getting non-cancer cases labeled as cancer-positive, thereby reducing stress and medical interventions unnecessarily for the patients. Recall, which measures the model's ability to correctly identify true positive instances, is complementary to Precision. This would be of paramount importance in the maximization of the detection of lung cancer cases. High recall rate minimizes the chance of missing positive diagnoses and thus helps in prompting clinical interventions that lead to early treatment.

The F1 Score is the harmonic mean of precision and recall, meaning it balances the model's ability to minimize both false positives as well as false negatives. This is useful for class-imbalanced contexts, where precision and recall can be folded into one number that represents a better comprehensive measure of performance. The last of which applied was the Cross-Validation Score, utilizing a 5-fold cross-validation to test its robustness and generalizability on how it could stand across different splits of the data. Cross-validation performs a test over multiple splits; such tests result in consistent results, thus reducing overfitting to any configuration of the data.

## IV. Result

All learning machines were evaluated concerning their performances on key performance metrics including Precision, Recall, F1-Score, Accuracy, and Cross-Validation Score. Table III summarizes the results concerning the relative performances of diverse classifiers for the prediction of lung cancer. The evaluation highlights the effectiveness of each classifier in capturing the nuances of the dataset, providing valuable insights into their suitability for accurate and reliable lung cancer prediction tasks.

Table III reports precision, recall, F1-score, accuracy, and cross-validation scores for each model taken into consideration. Classifier XGBoost the greatest accuracy; 96.64% having cross validation of 0.9434, demonstrating great predictive power and generalizing well to the observed levels of data splits. Other top-performing models include Support Vector Classifiers, Random Forest, Gradient Boosting, and Multi-layer Perceptron classifiers at an accuracy value of around 95.8% and high cross-validation scores.

The lowest accuracy, however, was realized by the Multinomial Naive Bayes classifier at 75.63% with a cross-validation score of 0.7542, indicating that the performance was not as effective with this data set.

TABLE III. Model Performance Summary

| Model | Precision | Recall | F-1 -Score | Accuracy | Cross-Validation |
|---|---|---|---|---|---|
| Logistic Regression | 0.94 | 0.94 | 0.94 | 0.9412 | 0.9034 |
| Decision Tree | 0.91 | 0.91 | 0.91 | 0.9076 | 0.9307 |
| KNN | 0.96 | 0.96 | 0.96 | 0.958 | 0.9099 |
| Gaussian Naive Bayes | 0.91 | 0.91 | 0.91 | 0.9076 | 0.8719 |
| Multinomial Naive Bayes | 0.76 | 0.76 | 0.76 | 0.7563 | 0.7542 |
| SVC | 0.96 | 0.96 | 0.96 | 0.958 | 0.9454 |
| Random Forest algorithm | 0.96 | 0.96 | 0.96 | 0.958 | 0.9392 |
| Gradient Boost | 0.95 | 0.95 | 0.95 | 0.9496 | 0.9434 |
| XGBoost | 0.97 | 0.97 | 0.97 | 0.9664 | 0.9434 |
| MLP Classifier | 0.96 | 0.96 | 0.96 | 0.958 | 0.9497 |
| Ensemble Voting Classifier | 0.96 | 0.96 | 0.96 | 0.958 | 0.941 |

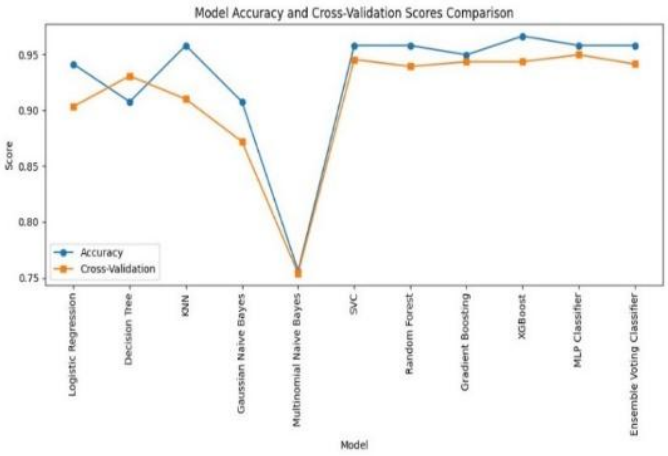


Fig. 2. Model Accuracy and Cross-Validation Scores Comparison

Figure 2: Line chart of accuracy (blue line) and cross-validation scores (orange line) for each model. Here, in this figure, it can be seen that most of the models have their accuracy and cross-validation scores in alignment, so these are generalizable models.

Once again, XGBoost and Voting Classifier have emerged as the top models based on both criteria. The enormous drop in the scores of the Multinomial Naive Bayes model visually emphasizes underperformance and further justifies the limits.

## V. Conclusion

The following shows the capability of machine learning algorithms to predict lung cancer risk using clinical and symptomatic data. Using a structured pipeline including data preprocessing, feature engineering, class balancing, and hyperparameter tuning, we evaluate ten individual machine learning models and an ensemble voting classifier. Among the

individual classifiers, XGBoost exhibited the highest performance with an accuracy of 96.64% and a cross-validation score of 0.9434. The Voting Classifier, which averaged the predictions of multiple models, was also found to be quite robust, with an accuracy of 95.8% and a cross-validation score of 0.9413.Such potential in early lung cancer detection applies to the clinical applications of both XGBoost and Voting Classifier models, as both achieved very high accuracy, precision, recall, and F1 scores. The data used for the study came solely from clinical and symptomatic observations.

Future research could potentially expand on this study by including those sources of data, and further enhancing the accuracy of the model can be done using ensemble techniques. Even the deeper architecture of learning might provide possible improvements for the models. Further refinement and tuning of the ML models should focus on increasing their robustness, ensuring transparency, and incorporating additional data types that better represent the full spectrum of lung cancer risk factors. The integration of multi-omics data, imaging, and clinical data will present opportunities for more accurate and holistic predictive models. With continuous advancements, machine learning models will play an increasingly critical role in improving the early detection of the patients possesing lungs cancera.

In general, this research study confirms that machine learning, particularly ensemble and gradient-boosting approaches, can indeed provide a viable, scalable solution predicting the lungs cancer. This is made possible because these methods enable healthcare practitioners to use a low-cost, non-invasive tool for the support of early diagnosis of cancer, thus facilitating early intervention in patients and improved outcomes.

## REFERENCES

[1] A. Y. Saleh, C. K. Chin, and R. A. Rosdi, "Transfer learning for lung nodules classification with CNN and random forest," Pertanika J. Sci. Technol., vol. 32, no. 1, pp. 463–479, 2024.

[2] M. V. Anand, et al., "Gaussian Naïve Bayes algorithm: A reliable technique involved in the assortment of the segregation in cancer," Mobile Inf. Syst., vol. 2022, Art. no. 2436946, 2022.

[3] C. S. Anita, et al., "Lung cancer prediction model using machine learning techniques," Int. J. Health Sci., vol. 6, no. S2, pp. 12533–12539, 2022.

[4] H. Beg, "Early detection of lung cancer using logistic regression algorithm," ResearchGate, 2022.

[5] P. Bhuvaneswari and A. B. Therese, "Detection of cancer in lung with K-NN classification using genetic algorithm," Procedia Mater. Sci., vol. 10, pp. 433–440, 2015.

[6] M. S. Bhuiyan, et al., "Advancements in early detection of lung cancer in public health: A comprehensive study utilizing machine learning algorithms and predictive models," J. Comput. Sci. Technol. Stud., vol. 6, no. 1, pp. 113–121, 2024.

[7] M. S. Bhat, "Lung cancer dataset," Kaggle, 2021. [Online]. Available: https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer.

[8] D. V. C. Lima, et al., "An integrated data analysis using bioinformatics and random forest to predict prognosis of patients with squamous cell lung cancer," IEEE Access, vol. 12, pp. 59335–59345, 2024.

[9] S. U. Krishna, et al., "Lung cancer prediction and classification using decision tree and VGG16 convolutional neural networks," Open Biomed. Eng. J., vol. 18, 2024.

[10] Lung cancer detection using combination of Gabor filter, histogram equalization and multi-layer perceptron," IEEE Int. Conf. Autom. Control Intell. Syst., vol. 2024, pp. 1979-8-3503-7210-6/24, 2024.

[11] 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. 59880–59899, 2024.

[12] P. Nanglia, et al., "A hybrid algorithm for lung cancer classification using SVM and neural networks," ICT Express, vol. 7, pp. 335–341, 2021.

[13] S. O. Olawale-Shosanya, et al., "A meta-ensemble predictive model for the risk of lung cancer," Al-Bahir J. Eng. Pure Sci., vol. 5, no. 1, pp. 1–10, 2024.

[14] P. L. Paelongan and I. Palupi, "Lung cancer prediction model using logistic linear regression with imbalanced dataset," Ind. J. Comput., vol. 7, no. 2, pp. 1–14, Aug. 2022.

[15] R. G. E. Patil, et al., "Lung cancer prediction system using logistic regression approach," Int. Res. J. Mod. Eng. Technol. Sci., vol. 2, no. 12, pp. 656–661, Dec. 2020.

[16] V. P. Patil, et al., "Design and development of lung cancer prediction model for performance enhancement using boosting ensembled machine learning classifiers with shuffle-split cross validations," J. Electron. Syst., vol. 20, pp. 9–28, 2024.

[17] Project code for lung cancer prediction using logistic regression," 2022. [Online]. Available: https://www.researchgate.net/.

[18] R. T. Noviandy, et al., "Machine learning approach to predict AXL kinase inhibitor activity for cancer drug discovery using Bayesian optimization-XGBoost," J. Soft Comput. Data Min., vol. 15, no. 1, pp. 46–56, 2024.

[19] S. Zhang, et al., "Predicting the risk of lung cancer using machine learning: A large study based on UK Biobank," Medicine, vol. 103, no. 16, Art. no. e37879, 2024.

[20] M. Vedaraj, et al., "Early prediction of lung cancer using Gaussian naive Bayes classification algorithm," Int. J. Intell. Syst. Appl. Eng., vol. 11, no. 6s, pp. 838–848, 2023.

[21] R. Vešović, et al., "A machine learning multilayer meta-model for prediction of postoperative lung function in lung cancer patients," Appl. Sci., vol. 14, no. 4, Art. no. 1566, Feb. 2024.

[22] X. Wang and M. Wang, "Classification of lung cancer with microRNA expression data using decision tree-based models," IEEE Trans. Biomed. Eng., vol. 67, pp. 1314–1322, 2019.

[23] Z. Ye, et al., "A naive Bayes model on lung adenocarcinoma projection based on tumor microenvironment and weighted gene co-expression network analysis," Infect. Dis. Model., vol. 7, pp. 498–509, 2022.

[24] Z. Zhou, et al., "A machine learning model for predicting lung cancer recurrence using clinical and molecular data," Front. Oncol., vol. 10, Art. no. 1150, 2020.

[25] H. Zhu, et al., "Lung cancer diagnosis with multi-class SVM and CT imaging," Oncol. Lett., vol. 18, pp. 2125–2131, 2019.

[26] Nguyen, H.T., Nguyen, M.N., Pham, S.C. et al. Abnormalities detection on chest radiograph with bounding box-based lungs extraction and object detection algorithm. Int. j. inf. tecnol. 16, 2241–2251 (2024). https://doi.org/10.1007/s41870-023-01687-9.

[27] Mittal, K., Aggarwal, G. & Mahajan, P. Performance study of K-nearest neighbor classifier and K-means clustering for predicting the diagnostic accuracy. Int. j. inf. tecnol. 11, 535–540 2019.

[28] Sahu, S.K., Kumar, P. & Singh, A.P. Modified K-NN algorithm for classification problems with improved accuracy. Int. j. inf. tecnol. 10, 65–70 (2018). https://doi.org/10.1007/s41870-017-0058-z.

[29] Bhardwaj, P., Kaur, A. Impact of image enhancement methods on lung disease diagnosis using x-ray images. Int. j. inf. tecnol. 15, 3521–3526 (2023). https://doi.org/10.1007/s41870-023-01409-1.

[30] Bhat, A.R., Hashmy, R. Hierarchical autoencoder-based multi-omics subtyping and prognosis prediction framework for lung adenocarcinoma. Int. j. inf. tecnol. 15, 2541–2549 (2023). https://doi.org/10.1007/s41870-023-01310-x.

[31] Rani, V.J., K.Thanammal, K. Lung cancer segmentation using MIBFS clustering and energetic BPN. Int. j. inf. tecnol. 15, 905–916 (2023). https://doi.org/10.1007/s41870-023-01164-3.

[32] Agrawal, S., Chowdhary, A., Agarwala, S. et al. Content-based medical image retrieval system for lung diseases using deep CNNs. Int. j. inf. tecnol. 14, 3619–3627 (2022). https://doi.org/10.1007/s41870-022-010077.

 