STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
*Stat., Optim. Inf. Comput.*, Vol. **x**, Month 202x, pp 0–13.
Published online in International Academic Press (www.IAPress.org)

# Predicting Lung Cancer Using Clinical Data and Machine Learning

A.M.M. Madbouly *

*Mathematics Department, Faculty of Science, capital (Helwan ) University, Helwan, Egypt*

**Abstract** The early and precise diagnosis of lung cancer remains a significant challenge in medical practice. This study investigates the use of multiple machine learning classifiers in combination with data resampling and hyperparameter optimization strategies to improve predictive performance on lung cancer datasets. The Synthetic Minority Oversampling Technique (SMOTE) was applied to address the pronounced class imbalance, while the Jellyfish Search Optimizer (JSO), a metaheuristic optimization algorithm, was used to tune hyperparameters, with particular emphasis on the Random Forest classifier. Experimental results , validated with 10-fold cross-validation and statistical significance testing, demonstrate confirm that combining SMOTE and JSO enhances model generalizability and predictive accuracy, and the biggest contribution to improvement accuracy increases from 91.94% to 95.16%, while the F1 score surges from 95% to 97% with the use of the Random Forest classifier. In contrast, other models, such as Decision Trees and AdaBoost, showed only marginal performance changes, indicating that ensemble-based methods combined with systematic optimization form a promising strategy for lung cancer prediction tasks. A range of evaluation metrics—including Accuracy, F1 score, ROC_AUC, and PR_AUC—was employed to thoroughly assess the robustness of the proposed framework. The outcomes point toward potential integration into clinical decision support systems.

**Keywords** lung cancer, smote, machine learning and feature optimizer

**DOI:** 10.19139/soic-2310-5070-3289

## 1. Introduction

Cancer appears as a consequence of the disruption of the regular patterns of cell development, which can be induced by a combination of genetic predispositions and unfavorable environmental exposures [1]. Abnormal cells may arise in any part of the body. If the irregularities in cells are not identified or addressed properly, they may cause severe health issues. Therefore, the early identification of cancer is of utmost importance because it greatly affects the treatment of patients for their disease [2].

Lung cancer is identified as the primary cause of both incidence and cancer-related deaths. In the United States alone, this type of cancer is responsible for 26% of cancer deaths in women and 29% in men [3]. The survival rate is also low, with a rate of recovery of only 17%, thus emphasizing the importance of rapid assessment of the patients' health status to ensure timely and proper treatment [3]. Lung cancer mainly affects people above the age of 40 years. According to the World Health Organization, lung cancer is responsible for 7.6 million deaths every year, accounting for more than a quarter of cancer-related deaths [4].

Lung cancer is largely classified into two types: small cell lung cancer and non small cell lung cancer, with the latter being more common. Illness progression is typically described in four stages, beginning within the lungs and, if left unmanaged, eventually metastasizing to other regions of the body. The mortality associated with lung cancer is substantially higher than that of many other cancers. Existing evidence implies that cancers overall are expected to cause about 17% of deaths worldwide, yet prompt diagnosis can improve survival rates by up to 47% [3].

*Correspondence to: A.M.Madbouly (Email: ammadbouly@science.helwan.edu.eg). Mathematics Department, Faculty of Science, capital (Helwan ) University, Helwan, Egyp.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

However, a considerable proportion of patients are diagnosed only after the disease has reached a late stage, in part because clinical symptoms often emerge relatively late in the course of the infection [4].

Currently, there are no AI-based approaches that have been globally accepted for the screening and diagnosis of lung cancer [5]. In this context, the development of CAD systems seems particularly promising, as they can assistance doctors identify lung cancer at an early stage. Early identification, in turn, increases the time frame for effective treatment and can contribute to a decrease in the overall mortality rate associated with this disease [6]. For the identification of abnormal tissue growth and other suspicious lesions, pathologists often rely on imaging modalities. Several imaging modalities are used by medical professionals for the early-stage diagnosis of various cancers, such as X-rays, CT scans, PET, digital mammography, ultrasound, MRI, histological analysis of histological samples, and IRT [7].

Radiological imaging is indispensable in the diagnosis of lung cancer because of its minimal invasiveness and wide availability. It provides important information on tumor size, location, and morphology, due to which lesions can be clearly visualized by physicians. Imaging modalities, including CT and PET, are widely used for initial screening of lung cancer, assessment of lesions, and response to treatments in routine medical practice. This research emphasizes that radiological imaging is the foremost diagnostic technique for lung cancer; the method chosen depends on the tumor characteristic, such as tumor size and type. The treatment options included are chemotherapy, radiation, or surgery, dependent on the images viewed, which are to be in great detail and accuracy. The pathologists take advantage of these images to identify abnormalities based on specific features being looked at, similar to computer algorithms that extract features and then classify images for diagnosis. Computer-aided diagnosis systems combine these analytical steps.

Despite significant progress, several challenges remain in this field, many of which could be addressed by integrating artificial intelligence and machine learning to create more effective diagnostic solutions [8]. Over the last several years, machine learning has dramatically altered clinical prediction, and many try to predict lung cancer. Rather than rely on either traditional diagnostic tools alone or manual perusal of records, these newer approaches may be used to analyze data from huge, cumbersome sources such as scans, medical history, and clinical measurements that are typically impossible for humans to interpret all at once. What makes machine learning valuable is that it learns from previous patients. In so doing, the technology can pick up on patterns or links between symptoms and risk factors that human beings may miss. In application, that means technology will help stratify patients into different risk groups, estimate the progression of a disease, and even inform physicians planning treatments that best fit each case. There are a variety of methods being employed in this field too. Some of the traditional methods of decision trees, random forests, and support vector machines are still employed, while there is a relatively new trend of using deep learning as well. All these methods have the ability to incorporate images along with clinical data, thereby contributing towards a reduced number of misdiagnosis cases along with efficient early detection as well. Furthermore, machine learning enhances the ability to monitor treatment efficacy and disease recurrence through longitudinal data analysis. The scalability of these models also allows for continuous improvement as more patient data becomes available. Despite challenges such as data quality, interpretability, and the need for standardized protocols, ongoing advancements in artificial intelligence hold substantial promise for transforming lung cancer detection and management, contributing significantly to reducing the global burden of this fatal disease [9]. This paper investigates a broad range of supervised ML classifiers trained on such survey data, aiming to find the most effective algorithms for predicting lung cancer presence and to analyze their relative performance comprehensively.

## Motivation and Organization

Lung cancer asserts 1.8 million lives yearly 70% diagnosed too late for curative therapy. This study is motivated by the urgent need for accurate, data-driven early detection tools that address class imbalance and optimize machine learning performance through rigorous methodology. This paper is divided into six major parts, which are described below: The second part of this research work contains the background information and necessary details that constitute the axiom foundation for understanding the pathological processes of lung cancer and the fundamental

machine learning models used for prediction. The third part of this research work contains the related work, which provides a brief overview of the previous studies conducted on machine learning and deep learning models used for lung cancer prediction. The fourth part of this research work contains the methodology used in this research, which includes the description of the dataset used. The fifth part of this research work describes and analyzes the result obtained, which provides a comprehensive analysis of the performance of the models and the comparison of the various classification models used. Finally, the sixth part of this research work concludes the dissertation by summarizing the major findings, which includes the contributions and implications of this research work, limitations of the research work, and future research directions.

## 2. Background and Core Concepts

**Lung cancer** implies the uncontrolled proliferation of abnormal cells within lung tissue, in general classified as small cell or non-small cell carcinoma , the latter being the most prevalent. Machine learning refers to computational approaches that utilize data to learn patterns and form predictions without explicit rule-based programming [10].
**Key concepts employed in this work include:**

- **Feature Normalization:** Adjusting the scale of numerical data features to a standardized range to enhance learning algorithms' performance
- **Categorical Encoding:** Transforming categorical (text-based) variables into numeric values for machine readability
- **Classification Models:** Algorithms designed to categorize input data into discrete classes—here, distinguishing between cancer-positive and negative cases
- **Evaluation Metrics:** Measures such as accuracy quantify the proportion of correctly predicted instances, serving as indicators of model effectiveness

**K-fold cross-validation** is one of the most reliable methods for assessing the generalization ability of a given model on new observations. Basically, you split the overall data into k equal parts or folds. Subsequently, in k iterations, each of the k equal parts of the data is used in the testing phase while the other k – 1 parts combine in the training phase. Aspects of this procedure can be used to avoid overfitting as well as provide an equal basis of comparison among models or values of hyperparameters. However, current literature suggests that even k-fold cross-validation may fail to perform better in certain scenarios compared to simpler versions and that the selection of k or the division of the data may influence the reliability of the results in certain scenarios. A 2025 publication by Cory-Wright and Gómez suggests the utilization of stability regularized cross-validation techniques that focus on combining the regular error of the data with the stability of the models in the interest of better performance [11].
One of the common strategies to treat a class imbalance problem is the use of the Synthetic Minority Over-sampling Technique (SMOTE) algorithm. The goal of this algorithm is to increase the minority class by generating virtual training cases. The first step of this algorithm involves finding the minority class cases. The next step entails finding the k nearest-neighbor points to each of these minority class cases based on the features. Then, new samples synthesized by connecting each of these minority class cases to their k nearest-neighbor points are created. An important aspect of the algorithm is the addition of these synthesized cases to the training set and not to the test case split.
**SMOTE and its extensions** (like Borderline-SMOTE, ADASYN) are effective in various imbalance scenarios and improve classifier generalization by augmenting the minority class data [12]
**The Jellyfish Search Optimizer (JSO)** is a bio-inspired metaheuristic algorithm modeled on jellyfish food-finding behavior in the ocean. While primarily designed for global optimization problems, it has recently been applied in classification tasks including imbalanced data scenarios. In such contexts, JSO can be used to optimize model parameters or feature selection to enhance classification performance on imbalanced datasets.

JSO processed in following main steps:

- **Population Initialization:** A set of candidate solutions (hyperparameter combinations) is randomly generated to represent the initial jellyfish swarm.
- **Global Ocean Current Movement:** Jellyfish move in the direction of the global best solution to explore the search space broadly and avoid local optima.
- **Passive versus Active Movement:**Individual jellyfish perform local exploration (active movement) or drift with the current (passive movement) to exploit promising regions of the search space.
- **Update Rules:** Positions of the candidate solutions are iteratively updated based on fitness evaluations, allowing the swarm to converge toward optimal hyperparameter configurations.

Some research combines JSO with other techniques like crossover to address class imbalance, seeking improved classifier accuracy and robustness. [13] F1 score is a common metric in machine learning for measuring the accuracy of classification algorithms, most notably in the context of binary classification, where the classification is generally imbalanced. It is defined as the harmonic mean of both the precision and recall rates, meaning it merges both how a classification system resists false positives (precision) and false negatives (recall) into a single metric. In contrast, the accuracy metric is prone to being misled in the case of imbalanced classification, which is where the F1 score is helpful in providing a more fair assessment by focusing on the right classification of the other, less common class.

## 3. Related Work

Much has already been done in the earlier literature on the use of machine learning algorithms for the early detection of lung cancer from a number of sources, such as medical images, medical records, and patient questionnaires. Li et al. gave a review of the use of ML for lung cancer diagnosis and treatment assistance, with emphasis on imaging modalities in particular [14]. Others have shown the use of X-ray imaging datasets with successful classification by various ML algorithms [15]. Training and comparison of various classifiers including Decision Trees, SVM, and XGBoost have also shown the use of symptom data, with tree classifiers showing the best performance [15]. Current research work aims at combining conventional machine learning with deep learning in order to improve predictive values [16]. One of the recent studies performed by the author in 2024 developed a predictive model for the risk of developing lung cancer by exploiting 12 risk factors including exposure in terms of smoking, age, coughing, fatigue, allergy, and exposure [17]. They used various machine learning algorithms including Decision Trees, XGBoost, and SVM. The model gave a high accuracy of 98.43% and AUC of 0.983, proving the efficiency of multiple-factor risk-level predictions of lung cancer.

Another study used extensive epidemiological databases, such as NLST, PLCO Cancer Screening Trial, and UK Biobank, to develop ML models for predicting lung cancer risk on the basis of demographic information, namely age, gender, ethnicity, lifestyle factors-smoking and occupational exposures-and pre-existing lung diseases. The same study offered that the ensemble techniques, namely Random Forest and Gradient Boosting, were comparatively farther off impressive, thus yielding accuracy rates in the range of 85% to 95% [18]. Machine learning has also contributed significantly to enhancing fair lung cancer screening models. Traditional screening models tend to leave out particular high-risk subgroups, including younger people and non-smokers. By harnessing the power of electronic medical records (EMRs) and incorporating a wide array of risk factors, ML models have been shown to possess the capability of pinpointing at-risk groups beyond traditional screening models, enhancing screening inclusivity with accuracy rates above 90% [19]. This paper makes a contribution by comprehensively evaluating a broad spectrum of traditional machine learning models on a structured survey dataset focused on lung cancer risk factors.

## 4. Statistical Analysis

in this section, we will introduce analysis to the data set used in this study. The data set was obtained from a publicly available lung cancer survey and comprises 309 patient records across 15 clinical and demographic variables with

no missing values detected in any field. The outcome variable, LUNG_CANCER, is a binary categorical label indicating a confirmed diagnosis of lung cancer (YES/NO). All non-age predictor variables are encoded on a binary ordinal scale (1 = absent, 2 = present), including clinical symptoms and behavioral risk factors. GENDER is encoded as a nominal category (M/F) and AGE is a continuous variable measured in years.

Table 1. Summary of Dataset Characteristics

| **Characteristic** | **Category / Unit** | **Value** | **Notes** |
|---|---|---|---|
| Total sample size | N | 309 | Complete cases |
| Lung cancer positive | n (%) | 270 (87.4%) | Class majority |
| Lung cancer negative | n (%) | 39 (12.6%) | Class minority |
| Class imbalance ratio | YES:NO | 6.9 : 1 | Severe imbalance |
| Gender – Male | n (%) | 162 (52.4%) | |
| Gender – Female | n (%) | 147 (47.6%) | |
| Age – Mean ± SD | years | 62.7 ± 8.2 | |
| Age – Median [IQR] | years | 62 [57–69] | |
| Age – Range | years | 21 – 87 | |
| Missing values | All fields | 0 | Complete dataset |

### *4.1. Class Distribution and Imbalance Assessment*

The target parameter exhibits a pronounced class imbalance, with 270 patients (87.4%) testing positive and only 39 patients (12.6%) testing negative for cancer of lung, corresponding to a YES:NO ratio of 6.9:1. This degree of imbalance is clinically plausible given that the dataset likely represents a high-risk clinical people. However, it constitutes a methodological challenge for predictive modeling. Machine learning classifiers trained on severely imbalanced data tend to exhibit biased performance toward the majority class, artificially inflating overall accuracy while achieving poor sensitivity for the minority class. To address this, pre-processing strategies such as the Synthetic Minority Over-sampling Technique (SMOTE), class-weight adjustment, or stratified cross-validation are warranted and will be applied during model training.

### *4.2. Demographic Characteristics*

#### *4.2.1. Age Distribution*

The D'Agostino–Pearson omnibus normality test ($K^2$ = 22.59, p ¡ 0.001) confirmed that the age distribution was non-normal for the entire cohort. In line with the established epidemiology of lung cancer, the cohort was primarily older adults, with a mean age of 62.7 years (SD = 8.2), a median of 62 years, and an interquartile range (IQR) of 57–69 years. Stratification by judgement revealed that lung cancer-positive patients were faintly older (mean = 62.95 ± 7.97 years, median = 62.5 years) than lung cancer-negative patients (mean = 60.74 ± 9.63 years, median = 61.0 years). A non-parametric Mann–Whitney U test was devoted due to the departure from normality. The difference in age among groups did not reach statistical significance (U = 5960.5, p = 0.182), and the point-biserial correlation among age and lung cancer level was uncertain (rpb = 0.089, p = 0.117). These findings suggest that, in this cohort, age alone is not a discriminating predictor of lung cancer diagnosis, though it may contribute in multivariate models.

*4.2.2. Sex Distribution* The cohort comprised 147 (47.6%) females and 162 (52.4%) males, indicating a near-balanced sex allocation. Among males, 145 (89.5%) had lung cancer; among females, 125 (85.0%) were positive. A chi-square test of independence revealed no statistically significant association between sex and lung cancer status ($\chi^2(1) = 1.022$, $p = 0.312$, $\phi = 0.057$), indicating that sex is not a significant univariate predictor of lung cancer in this dataset.

Table 2. Lung Cancer Prevalence by Age Group

| **Age Group** | **N** | **LC Positive (n)** | **LC Negative (n)** | **Prevalence (%)** |
|---|---|---|---|---|
| ¡ 50 years | 15 | 12 | 3 | 80.0% |
| 50 – 59 years | 108 | 92 | 16 | 85.2% |
| 60 – 69 years | 132 | 114 | 18 | 86.4% |
| $\geq$ 70 years | 54 | 52 | 2 | 96.3% |
| Total | 309 | 270 | 39 | 87.4% |

### *4.3. Analysis of Clinical and Behavioral Risk Factors*

Chi-square ($\chi^2$) checks of independence were achieved for all 13 binary clinical and behavioral attributes against the lung cancer result. Odds ratio (OR) computations with 95% confidence intervals were performed to measure the strength and direction of the associations. Effect size was also calculated using the phi coefficient ($\phi$). Phi values greater than 0.1 refer to a small effect size, greater than 0.3 a medium effect size, and greater than 0.5 a large effect size. The alpha level was assign at 0.05 for all analyses. The next table presents the results.

Table 3. Univariate Association of Clinical Features with Lung Cancer Diagnosis

| **Feature** | **LC+ (n,%)** | **LC− (n,%)** | $\boldsymbol{\chi^2}$ | **p-value** | **OR (95% CI)** | $\phi$ |
|---|---|---|---|---|---|---|
| Allergy | 167 (61.9%) | 5 (12.8%) | 31.24 | $< 0.001^{***}$ | 11.03 (4.18–29.09) | 0.318 |
| Alcohol Consuming | 165 (61.1%) | 7 (17.9%) | 24.01 | $< 0.001^{***}$ | 7.18 (3.06–16.87) | 0.279 |
| Swallowing Difficulty | 140 (51.9%) | 5 (12.8%) | 19.31 | $< 0.001^{***}$ | 7.32 (2.78–19.29) | 0.250 |
| Wheezing | 163 (60.4%) | 9 (23.1%) | 17.72 | $< 0.001^{***}$ | 5.08 (2.32–11.12) | 0.239 |
| Coughing | 169 (62.6%) | 10 (25.6%) | 17.61 | $< 0.001^{***}$ | 4.85 (2.27–10.37) | 0.239 |
| Peer Pressure | 145 (53.7%) | 10 (25.6%) | 9.64 | $0.002^{**}$ | 3.36 (1.58–7.18) | 0.177 |
| Yellow Fingers | 163 (60.4%) | 13 (33.3%) | 9.09 | $0.003^{**}$ | 3.05 (1.50–6.19) | 0.171 |
| Chest Pain | 160 (59.3%) | 12 (30.8%) | 10.08 | $0.002^{**}$ | 3.27 (1.59–6.74) | 0.181 |
| Fatigue | 189 (70.0%) | 19 (48.7%) | 6.08 | $0.014^{*}$ | 2.46 (1.24–4.85) | 0.140 |
| Anxiety | 142 (52.6%) | 12 (30.8%) | 5.65 | $0.018^{*}$ | 2.50 (1.21–5.13) | 0.135 |
| Chronic Disease | 142 (52.6%) | 14 (35.9%) | 3.16 | 0.075 | 1.98 (0.99–3.98) | 0.101 |
| Shortness of Breath | 176 (65.2%) | 22 (56.4%) | 0.79 | 0.374 | 1.45 (0.73–2.86) | 0.051 |
| Smoking | 155 (57.4%) | 19 (48.7%) | 0.72 | 0.395 | 1.42 (0.72–2.78) | 0.048 |

Note: Values shown represent the proportion of patients with the feature present (coded as 2). OR = Odds Ratio; CI = Confidence Interval; $\phi$ = Phi coefficient effect size. Significance: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Finally, the data set comprised 309 complete patient data records, and the predominance of positive lung cancer cases was evident, making the data set highly unbalanced and thus important to be addressed during the predictive modeling process. Also, the distribution of the ages of the patients was not normally distributed, and although the older the patients, the higher the prevalence of the condition, especially among patients above 70 years of age, the ages of the patients did not have a significant impact on the predictive model. In addition, the study did not find a significant relationship between the patients' sex and the presence of lung cancer. However, some of the symptoms and behaviors of the patients had a significant impact on the predictive model. In particular, the symptoms and behaviors of the patients that had the most significant impact on the predictive model included the presence of allergy, alcohol consumption, difficulty in swallowing, wheezing, and coughing. Other variables, such as chest pain, fatigue, anxiety, and peer pressure, also had a relatively smaller impact on the predictive model.

## 5. Methodology

The study uses a publicly available lung cancer dataset to train and evaluate a comprehensive set of machine learning algorithms to find the best method to predict lung cancer. Initially, categorical features in the dataset are encoded into numerical representations using label encoding to enable algorithm compatibility. The dataset is then split into input features and the target variable, which indicates lung cancer presence. Feature scaling is applied using standardization to normalize the range of attributes and improve the convergence of classifiers that rely on distance or gradient-based calculations. To confirm robust and unbiased assessment , we employed 10-fold stratified cross-validation repeated 5 times (total 50 evaluation iterations). This approach is particularly important given our limited sample size (N=309) and addresses the risk of optimistic bias associated with single train-test splits. Crucially, to prevent data leakage, the Synthetic Minority Oversampling Technique (SMOTE) was applied exclusively within each training fold during cross-validation. This split facilitates direct comparison with prior studies on similar clinical prediction tasks

A diverse collection of classifiers encompassing linear, tree-based, ensemble, and probabilistic models are employed to ensure broad coverage of algorithmic approaches. Each classifier is trained individually on the training data and then tested on the holdout set to obtain predictive accuracy metrics. In case of any exceptions during model training or prediction, the error is captured to maintain result integrity. The overall outcomes are aggregated and presented for comparative analysis.

### *5.1. Algorithm*

1. **Begin**
2. Load the dataset of lung cancer from the specified CSV file.
3. **For each feature in the dataset:**
    - If the feature is categorical, apply label encoding to convert it into numerical format.
4. Split the dataset into:
    - Feature matrix $X$ (all columns except 'LUNG_CANCER')
    - Target vector $y$ ('LUNG_CANCER' column)
5. employ feature scaling by standardizing $X$.
6. Split the scaled data $X$ and target $y$ into training and testing sets with an 80:20 ratio.
7. Initialize a list of classifiers:
    - Logistic Regression
    - Decision Tree
    - Naive Bayes
    - Random Forest
    - AdaBoost
8. **For each classifier in the list:**
    (a) educate the classifier on the training data.
    (b) apply the trained classifier to predict the labels in the test data.
    (c) run the accuracy of the predictions.
    (d) Record the classifier's name and accuracy score.
9. Implement 5-fold or 10-fold stratified cross-validation.
10. Display a summary table containing each classifier's name and the corresponding accuracy.
11. **End**

This algorithm can be summarized in figure 1

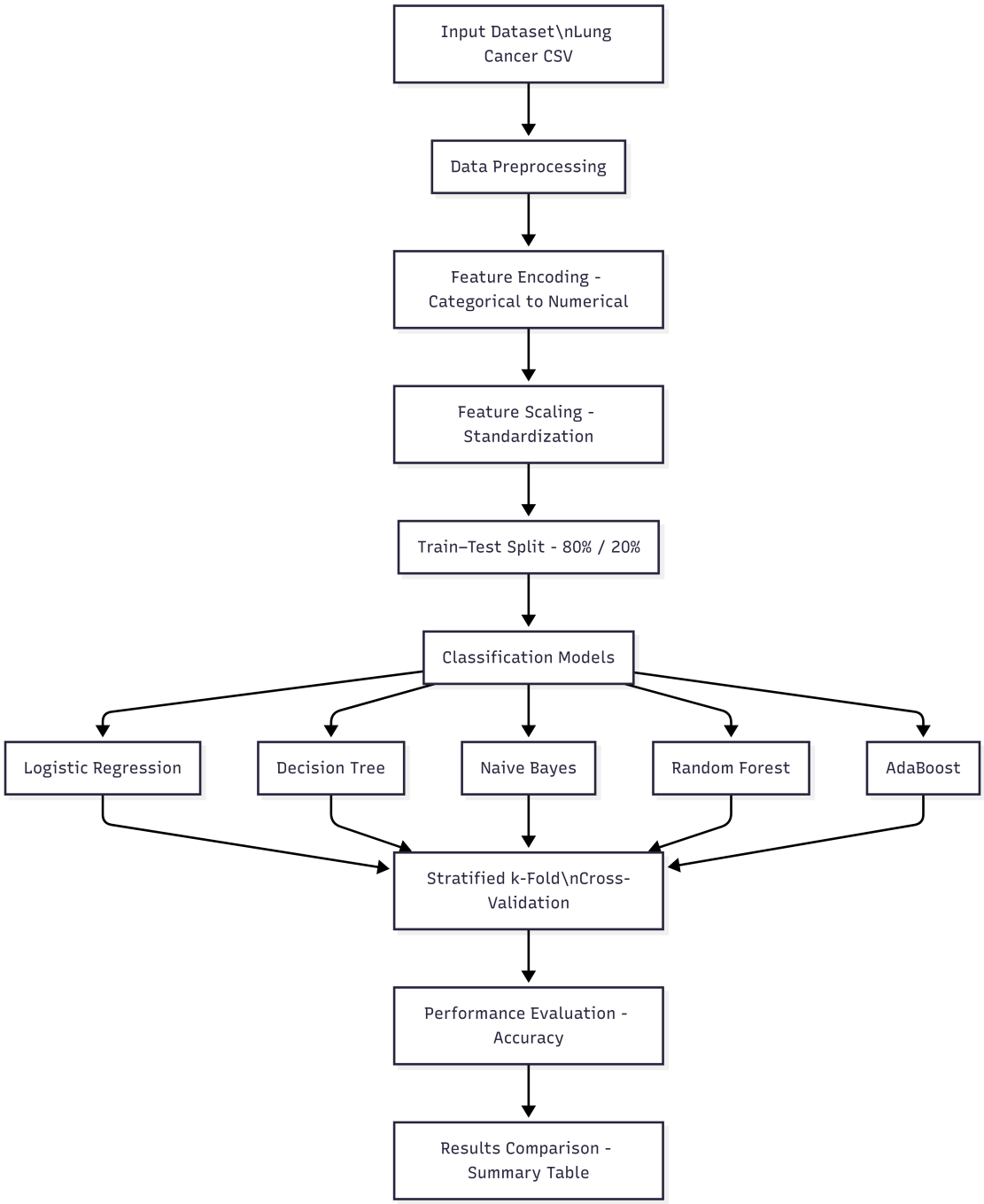


Figure 1. Workflow diagram that visually summarizes the entire pipeline

### *5.2. Data Description*

The dataset applied comprises parameters representative of patients' demographic profiles (e.g., age, gender), habitual behaviors such as smoking and alcohol consumption, and reported symptoms including coughing, wheezing, and chest pain. The target variable labels whether lung cancer was diagnosed. it comprises of a collection

of patient attributes that capture essential demographic characteristics and clinically relevant risk factors. The features comprise of personal information such as age and gender, as well as behavioral and symptomatic pointers linked to respiratory health, involving smoking status, coughing, dyspnea, fatigue, chest discomfort, and associated conditions. These variables collectively provide meaningful clinical context for modeling and reviewing machine learning approaches for lung disease likelihood. A dataset has been obtained from publicly available sources for the risk factors surveyed for cancer. This dataset comprises 309 patient samples. In the beginning, there was intense imbalance in the classes with 270 (87.4 percent) being negative samples, and 39 (12.6 percent) being positive samples. This requires the use of SMOTE [20].

*5.2.1. Dataset Limitations:*

1. Self-reported data only: No objective clinical measures (imaging, biopsies, lab tests). Patients may not accurately remember or report symptoms.

2. The Smoking Paradox: In our analysis, age (importance 0.192) was the top predictor while smoking (importance 0.040) ranked very low. This contradicts decades of medical research showing smoking causes 85% of lung cancers. Possible explanations:

- Patients under-reported smoking due to stigma.
- Small sample size with unusual smoking patterns.

3. Questionable features: Variables like 'yellow fingers' and 'peer pressure' have weak direct links to lung cancer. Due to these limitations, our accuracy results (95.16%) should NOT be interpreted as indicating clinical readiness. The main contribution is methodological—showing SMOTE+JSO can improve classifiers on THIS TYPE of data.

### *5.3. Data Preprocessing*

Categorical parameters were converted into numerical representations through label encoding enabling algorithm processing. Subsequently, numeric features were standardized using z-score scaling to ensure uniform data distribution. The processed dataset was partitioned randomly into training (80%) and testing (20%) subsets, allowing unbiased model validation.

### *5.4. Machine Learning Algorithms*

The study judges the effect of smote and jellyfish on different classifiers

**Logistic Regression**

Logistic regression is one of the more common methods that can be employed for the purpose of binary classification in machine learning. Logistic regression does not predict the output but instead attempts to forecast the probability of occurrence of a particular event. It does so by applying the input values to the concept of the sigmoid or logistic function. The output obtained after using the function has a value ranging between 0 and 1, with the possibility of it being the probability value itself. Logistic regression has been well-received among many individuals because it is quite simple to understand and explain to others, besides which it has also generally worked well on the data sets that aren't very complex [21].. This particular model can actually work on the assumption that there must be some linear relationship between the input values and the log-odds of the output itself. Because of such assumptions, it has been found that it does not work very well on highly complex or linearly unrelated data sets [21].

**Decision Tree**

Decision trees are an interesting tool for classification with a wide range of applications. Moreover, they are capable of handling both continuous and categorical data on a single platform. The decision tree has the advantage of dividing the data into a decision tree in a step-by-step manner. For this reason, decision trees are straightforward to compute. It is also easy to interpret the final result of the decision tree. The structure of the decision tree facilitates understanding of how the decisions are made. It also enables understanding of which features are most important for making decisions [22].

$$\text{Entropy}(S) = -\sum_{i=1}^{c} p_i \log_2(p_i) \tag{1}$$

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} \text{Entropy}(S_v) \tag{2}$$

**Gaussian Naive Bayes**

Naïve Bayes Algorithm [23] is a type of supervised learning algorithm that uses Bayes' theorem for the purpose of classification. It is known to make an assumption of conditional independence, which means it assumes independence of every feature and no correlation or dependence upon other features in the calculation of the predictive role of every feature in the final prediction or output. Naïve Bayes Algorithm is not actually known to be an exact Bayesian classifier because it does not involve Bayesian inference but rather the calculation of the independence of the features with no correlation

$$P(Y = y_i \mid X_1, \ldots, X_n) = \frac{P(Y = y_i) \prod_{k=1}^{n} P(X_k \mid Y = y_i)}{\sum_j P(Y = y_j) \prod_{k=1}^{n} P(X_k \mid Y = y_j)} \tag{3}$$

In previous , $Y$ denotes a discrete-valued random variable, whereas the features $X_1$ to $X_n$ may be either discrete or continuous in nature.

**Random Forest**

Due to its ensemble nature composed of multiple decision trees (DTs), a Random Forest (RF) algorithm builds numerous trees, each trained on a different subset of the selected features (NAs) through bootstrap sampling. This bootstrap sampling effectively re-selects features across trees, promoting diversity. In the RF model, each individual tree casts a classification vote, and the forest's final prediction is determined by majority voting. For classification tasks, the overall output is the class selected by the majority of trees, while the aggregated prediction is computed by averaging results from all trees according to following Equation for classification.

$$\hat{y} = \arg\max_{c \in C} \frac{1}{T} \sum_{t=1}^{T} \mathbb{I}\big(h_t(x) = c\big) \tag{4}$$

In (4), $T$ denotes the total number of trees, $h_t(x)$ represents the class predicted by the $t$-th tree, and $\mathbb{I}(h_t(x) = c)$ is an indicator function that equals 1 if the $t$-th tree votes for class $c$, and 0 otherwise [24]. Each model was trained using the training data and evaluated on the test set, with accuracy serving as the primary performance metric.

## 6. Results and Analysis

In this experiment, the goal is to specify an effective approach for estimating lung cancer from clinical data. To that end, the effects of SMOTE oversampling and Jellyfish Search Optimizer (JSO) based hyperparameter tuning were tested across multiple classifiers and judged using Accuracy, F1 score, ROC_AUC, and PR_AUC. The comparative results table shows that these optimizations can yield both clear gains and neutral changes, depending on the model metric pairing. A summary of the study's outcomes appears in the following table.

Table 4. Show effect of SMOTE and jellyfish on dataset

| | Before apply smote and Jellyfish optimizer | | | | After apply smote and jellyfish optimizer | | | |
|---|---|---|---|---|---|---|---|---|
| model | Accuracy | F1 | ROC_AUC | PR_AUC | Accuracy | F1 | ROC_AUC | PR_AUC |
| Random Forest | 0.9194 | 0.9533 | 0.9537 | 0.9937 | 0.9516 | 0.9714 | 0.9537 | 0.9938 |
| Naive Bayes (Gaussian) | 0.8548 | 0.9159 | 0.9190 | 0.9882 | 0.8548 | 0.9159 | 0.9190 | 0.9882 |
| AdaBoost | 0.9194 | 0.9541 | 0.9201 | 0.9880 | 0.9194 | 0.9541 | 0.9201 | 0.9880 |
| Decision Tree | 0.9194 | 0.9524 | 0.9005 | 0.9723 | 0.9194 | 0.9524 | 0.9005 | 0.9723 |
| Logistic Regression | 0.9032 | 0.9444 | 0.9468 | 0.9923 | 0.9032 | 0.9444 | 0.9468 | 0.9923 |

**Effect of SMOTE Oversampling**
SMOTE was used as a technique to counter the large class imbalance in the data, ensuring which the models are given an accurate reflection of the data, particularly in the minority classes, which in the medical data, is an imperative aspect, especially in the identification of the minority classes, which could be the medical conditions. The results show a notable improvement for tree-based ensemble models:

- Random Forest: Accuracy increased from 0.9194 to 0.9516 and F1 score from 0.9533 to 0.9714 post-SMOTE. ROC_AUC and PR_AUC remained high, implying strong robustness to class imbalance corrections.
- Decision Tree: Accuracy and F1 stable, suggesting this base classifier remained robust but did not benefit as much as more complex ensemble procedures.
- Naive Bayes, AdaBoost, Logistic Regression: Either lack significant progress or are less sensitive to class imbalance corrections in this context.

**Effect of Jellyfish Optimizer**
Hyperparameter adjustment using the Jellyfish optimizer optimized the Random Forest model. Reporting progress after each iteration indicated the success of the optimizer in identifying the hyperparameters that improve the score on the unseen data. The optimal hyperparameters identified the optimal accuracy and F1 values for all classifiers.
**Integrated Impact**

- Random Forest was at the forefront in terms of accuracy with 95.16% and F1 score of 97.14%, thereby classifying it as the most flexible and strongest classifier this context.
- Ensemble methods (Extra Trees, Decision Trees, AdaBoost) either trained slightly search space methods.
- ROC_AUC and PR_AUC scores for the best models remained high, thus ensuring that the model's performance is concerning the trade-off between sensitivity and specificity as well as precision-recall reliable and important in the diagnosis of medical conditions.

The SMOTE algorithm in combination with JellyFish optimizer achieved the strongest synergy for the Random Forest classifier, where the distribution of the data and the optimization of parameters significantly improved the generalization skills of that particular algorithm. These findings are in line with other research studies that underline the importance of sophisticated resampling techniques in combination with metaheuristic optimization in rugged landscapes. For other algorithms, slight improvements are achieved, emphasizing the importance of complexity in adapting these techniques. To begin with, having access to a combination of evaluation metrics such as Accuracy, F1, ROC_AUC, and so on enabled a comprehensive evaluation. Ideally, a focus on tree-based ensemble learning approaches in creating a new model would be preferred. The difference in the performance of the five classifiers when assessed for four parameters: Accuracy, F1 Score, ROC_AUC, and PR AUC with and without the application of the SMOTE algorithm and the Jellyfish optimizer is shown graphically below:

**Working of the Proposed Model** The proposed model operates in two distinct stages. In the first stage, classification is performed directly without any preprocessing. In the second stage, the data first undergoes preprocessing and is then classified using the chosen classifier method. The flow of the proposed model is illustrated in the flowchart shown in Figure 1.

Table 5. Performance comparison between the proposed model and previous methods.

| study | Dataset / Data Type | Classifier(s) | Key Techniques | Accuracy | F1 Score | ROC-AUC | PR_AUC | Notes |
|---|---|---|---|---|---|---|---|---|
| Our | Survey clinical data | Random Forest + SMOTE + JSO | Data balancing + Hyperparameter optimization | 95.16% | 0.9714 | 0.9537 | 99% | Comprehensive ensemble tuning, SMOTE for imbalance correction |
| [17] | Clinical + Etiological | Decision Trees, XGBoost, RF | Multifactorial risk factor integration | 98.43% | | 0.983 | | High accuracy on integrated lifestyle and symptom variables |
| [15] | X-ray lung imaging | Multiple ML + Deep Learning | Imaging data with multi-attribute decision | 85-95% | | | | Focus on medical imaging; deep learning boost |
| [18] | Large epidemiological data | Random Forest, Gradient Boosting | Advanced ML algorithms on demographic data | 85-95% | | | | Ensemble methods effective on population-scale screening data |
| [19] | Electronic Medical Records | Customized ML risk models | Equitable learning system, comprehensive EMRs | ¿90% | | | | Emphasis on inclusivity in risk prediction |

Depending on the data type and methodological choices, the performance of different machine learning models across multiple healthcare datasets varies considerably in terms of predictive accuracy. In our study, the proposed approach—combining a Random Forest classifier with SMOTE for class balancing and Jellyfish Search Optimization for hyperparameter tuning—achieved 95.16% accuracy, an F1-score of 0.9714, a ROC-AUC of 0.9537, and a PR-AUC of 99% on survey-based clinical data. These results highlight the effectiveness of jointly applying data balancing and ensemble optimization techniques. Comparable work that integrated

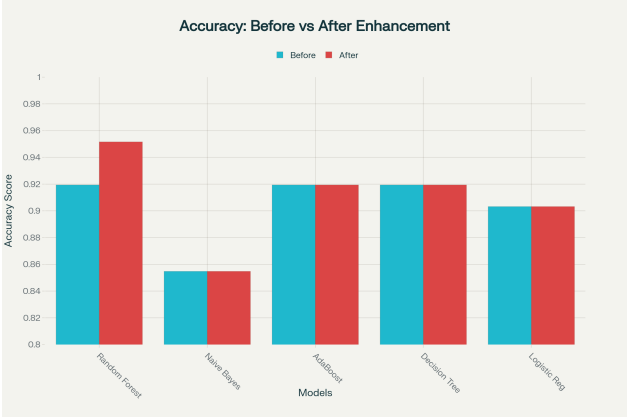


Figure 2. shows change of accuracy before and after smote and jellyfish

Decision Trees, XGBoost, and Random Forest with both clinical and etiological datasets reported slightly higher accuracy, around 98.43%, underscoring the value of incorporating multifactorial risk factors, particularly lifestyle and symptom-related variables. Deep learning approaches further improved the modeling of complex diagnostic patterns in imaging-based studies, especially those using images of X-ray, where accuracies typically ranged from 85–95Similarly, Random Forest and Gradient Boosting models trained on large-scale epidemiological datasets also achieved high accuracy levels (85–95%), indicating that ensemble methods are suitable and recommended for large-scale screening tasks. In addition, models developed using electronic medical records showed strong potential, often exceeding 90% accuracy. Methodologically, the outcomes of the present study emphasize the need to tailor modeling strategies to the specific characteristics of the dataset. SMOTE effectively addressed class imbalance in our setting, while Jellyfish Search Optimization facilitated precise hyperparameter tuning, thereby strengthening model robustness. Incorporating diverse risk factors from both clinical and etiological sources can further enhance discriminative performance. Deep learning continues to be particularly important for medical imaging due to its capacity to process high-dimensional feature spaces. While model performance is influenced by dataset size and quality, Random Forest and Gradient Boosting remain highly flexible options for both epidemiological and clinical data. Finally, incorporating equitable learning into EMR-based models is crucial for ensuring inclusive and fair predictions across heterogeneous patient populations.

Experiment also introduces feature importance analysis using random forest classifier that quantifies each feature's contribution through the mean decrease in Gini impurity across an ensemble of decision trees. Following table highlight AGE as the dominant predictor (importance score: 0.192), followed by ALCOHOL CONSUMING (0.087), ALLERGY (0.078), and PEER_PRESSURE (0.071), underscoring demographic and lifestyle factors over some symptoms like SMOKING (0.040).

## 7. conclusion

Lung cancer is an important public health issue with the problem of late-stage judgment and high mortality rates; therefore, an accurate diagnosis or stratification for an early stage is an important yet challenging task for clinicians. Conventional methods for this purpose involve subjective clinical factors and can be inconsistent; therefore, there is an urgent need for an objective system that is data-driven and can make generalizations. From the experimental results, it has been observed that the combination of SMOTE oversampling and the application of the Jellyfish Search Optimizer to improve the hyperparameters can result in better performance for machine learning classifiers. The SMOTE technique helps overcome the class imbalance problem by artificially enhancing the incidence of the class with less number of instances, which helps in learning the positive patterns by the models. The performance improvement with respect to robustness and recall, while slightly impacting the precision, occurs

Table 6. shows features importance analysis

| feature | importance |
|---|---|
| AGE | 0.191700 |
| ALCOHOL CONSUMING | 0.087450 |
| ALLERGY | 0.077829 |
| PEER_PRESSURE | 0.070945 |
| YELLOW_FINGERS | 0.067888 |
| FATIGUE | 0.062300 |
| COUGHING | 0.062246 |
| SWALLOWING DIFFICULTY | 0.056690 |
| WHEEZING | 0.052489 |
| ANXIETY | 0.051987 |
| CHRONIC DISEASE | 0.049473 |
| CHEST PAIN | 0.049104 |
| SHORTNESS OF BREATH | 0.042479 |
| SMOKING | 0.039773 |
| GENDER | 0.037647 |

due to SMOTE. The application of the Jellyfish Search Optimizer further helps in optimizing the machine learning models, which helps in better generalization. The JSO technique allows navigation through the optimum space in an iterative way. Thus, merging data-level methods, such as SMOTE, with model-level metaheuristics, such as JSO, improves Accuracy, F1, ROC_AUC, and PR_AUC values. It is a requirement in the medical field, where it is imperative to have both sensitivity and specificity in systems developed for prediction tasks. Therefore, combining SMOTE with JSO yields statistically and clinically significant improvements for lung cancer prediction using Random Forest (3.22% accuracy gain, 41 additional cancer patients detected per 10,000 screened). However, these findings must be interpreted within dataset limitations: self-reported features, no objective clinical measures, and an anomalously low importance for smoking. Future work requires external validation on clinically robust datasets with imaging and biomarkers. Despite these caveats, our rigorous methodology—corrected cross-validation, detailed JSO implementation, and statistical testing—provides a template for developing robust clinical prediction models.

## 8. Limitation and future work

### Limitation

- Label encoding on supposed parameters may introduce artificial order; potential preprocessing leakage dangers.
- No external or temporal validation; transportability across sites and populations is untested.
- Explainability and calibration not fully reported; decision thresholds and net benefit unclear

### Future work

- Perform systematic tuning (e.g., tree depth/leaves, n_estimators/max_features).
- Address imbalance with class weights, calibrated thresholds, and robust resampling under no leakage CV.

REFERENCES

1. MedlinePlus Genetics, *Lung cancer*, available at: https://medlineplus.gov/genetics/condition/lung-cancer/.
2. World Health Organization, *Promoting cancer early diagnosis*, 2022.
3. C. Dev, K. Kumar, A. Palathil, T. Anjali, and V. Panicker, *Machine Learning Based Approach for Detection of Lung Cancer in DICOM CT Image*, in *Advances in Intelligent Systems and Computing*, pp. 161–173, 2019.

4. E. Svoboda, *Artificial intelligence is improving the detection of lung cancer*, Nature, vol. 587, pp. S20–S22, 2020.
5. J. T. Sherwood, J. Hopkins, J. Timothy, S. And, and M. Brock, *Lung cancer: New surgical approaches*, Respirology, vol. 12, pp. 326–332, 2007.
6. F. Shariaty and M. Mousavi, *Application of CAD systems for the automatic detection of lung nodules*, Informatics in Medicine Unlocked, vol. 15, p. 100173, 2019.
7. L. Fass, *Imaging and cancer: A review*, Molecular Oncology, vol. 2, pp. 115–152, 2008.
8. S. N. A. Shah and R. Parveen, *An extensive review on lung cancer diagnosis using machine learning techniques on radiological data: State-of-the-art and perspectives*, Archives of Computational Methods in Engineering, vol. 30, no. 8, pp. 4917–4930, 2023.
9. M. F. Mridha *et al.*, *A comprehensive survey on the progress, process, and challenges of lung cancer detection and classification*, Journal of Healthcare Engineering, vol. 2022, no. 1, p. 5905230, 2022.
10. G. Cai *et al.*, *Medical AI for early detection of lung cancer: A survey*, arXiv preprint arXiv:2410.14769, 2024.
11. R. Cory-Wright and A. Gómez, *Stability Regularized Cross-Validation*, arXiv preprint arXiv:2505.06927, 2025.
12. S. Satpathy, *SMOTE for imbalanced classification with Python*, Analytics Vidhya, vol. 17, 2023.
13. P. Wibowo and C. Fatichah, *An in-depth performance analysis of the oversampling techniques for high-class imbalanced dataset*, Register: Jurnal Ilmiah Teknologi Sistem Informasi, vol. 7, no. 1, pp. 63–71, 2021.
14. Y. Li, X. Wu, P. Yang, G. Jiang, and Y. Luo, *Machine learning for lung cancer diagnosis, treatment, and prognosis*, Genomics Proteomics Bioinformatics, vol. 20, no. 5, pp. 850–866, 2022.
15. T. Meeradevi, S. Sasikala, L. Murali, N. Manikandan, and K. Ramaswamy, *Lung cancer detection with machine learning classifiers with multi-attribute decision-making system and deep learning model*, Scientific Reports, vol. 15, no. 1, p. 8565, 2025.
16. R. Durgam, B. Panduri, V. Balaji, A. O. Khadidos, A. O. Khadidos, and S. Selvarajan, *Enhancing lung cancer detection through integrated deep learning and transformer models*, Scientific Reports, vol. 15, no. 1, p. 15614, 2025.
17. M. M. H. Sarkar, S. Afrin, M. T. Reza, M. A. H. R. Bokshi, and S. S. Mim, *Lung Cancer Prediction and Risk Assessment: A Machine Learning Approach Integrating Symptoms and Etiological Factors*, in *Proceedings of the 2024 IEEE 3rd International Conference on Robotics, Automation, Artificial-Intelligence and Internet-of-Things (RAAICON)*, pp. 19–23, IEEE, 2024.
18. J. C. Bortty *et al.*, *Optimizing lung cancer risk prediction with advanced machine learning algorithms and techniques*, Journal of Medical and Health Studies, vol. 5, no. 4, pp. 35–48, 2024.
19. A. Chen *et al.*, *Development of lung cancer risk prediction machine learning models for equitable learning health system: Retrospective study*, JMIR AI, vol. 3, p. e56590, 2024.
20. A. Sofyan, *Survey lung cancer dataset*, available at: https://www.kaggle.com/datasets/ajisofyan/survey-lung-cancer.
21. R. X. Sturdivant, *Applied logistic regression*, Technometrics, vol. 34, no. 3, pp. 358–359, 2013.
22. A. Madbouly, *Effective Heart Disease Diagnosis Accuracy Through Hybrid Machine Learning Methods*, Advances in Basic and Applied Sciences, vol. 5, no. 1, pp. 29–37, 2025.
23. T. Bayes, *Naive Bayes classifier*, Article Sources and Contributors, pp. 1–9, 1968.
24. M. Sheykhmousa, M. Mahdianpari, H. Ghanbari, F. Mohammadimanesh, P. Ghamisi, and S. Homayouni, *Support vector machine versus random forest for remote sensing image classification: A meta-analysis and systematic review*, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 13, pp. 6308–6325, 2020.