# **Comparative Analysis of Machine Learning Models for Lung Cancer Detection**

**Prof. Pradeep Kumar K**

Assistant Professor, Dept. of CSE(AIML),

Dayananda Sagar University- Bengaluru [7661pradeep@gmail.com](mailto:7661pradeep@gmail.com)

**Mr. Chethan K Murthy**

2nd Year UG, Dept. of CSE(AIML),

Dayananda Sagar University- Bengaluru

[chethankeshavmurthy@gmail.com](mailto:chethankeshavmurthy@gmail.com)

# **Abstract:**

One of the most common and fatal types of cancer in the world today is lung cancer. Improving patient outcomes and survival rates is contingent upon early identification. We compare five machine learning models—Random Forest, Logistic Regression, Gaussian Naive Bayes, Bernoulli Naive Bayes, and Support Vector Machine (SVM)—for the purpose of identifying lung cancer in this work. A dataset comprising features taken from medical imaging scans and patient demographic data is used to train and assess the algorithms. Metrics including recall, F1-score, accuracy, and precision are used to assess performance. We also explore the models' interpretability in order to learn more about the characteristics that influence classification choices. Observed Results show that machine learning is effective in detecting lung cancer, albeit effectiveness varies depending on the model.

This study offers insightful information about the advantages and disadvantages of various machine learning techniques for lung cancer detection. Comprehending the performance attributes of these models is essential for making well-informed decisions in clinical practice and for creating trustworthy screening instruments for early identification and remediation.

Keywords: Random Forest, Logistic Regression, Gaussian Naive Bayes, Bernoulli Naive Bayes, and Support Vector Machine (SVM), accuracy, precision, recall, and F1-score

1. **Introduction:**

Being one of the primary causes of cancer-related death, lung cancer is a major global public health concern. Lung cancer patients continue to have a poor prognosis despite advancements in medical technology and treatment methods; this is mainly because of late-stage diagnosis and a lack of viable treatment alternatives. Because it allows for prompt intervention and raises the chance of a successful course of therapy, early identification of lung cancer is critical to bettering patient outcomes.

The application of machine learning techniques to the detection and diagnosis of lung cancer has garnered increasing attention in recent years. Machine learning algorithms have shown promise in the analysis of intricate medical data, such as patient demographics and imaging scans, to find trends and markers linked to the existence of lung cancer. These algorithms may be used in conjunction with already available diagnostic instruments to increase the efficacy and accuracy of lung cancer detection while also giving physicians insightful information.

In this regard, the purpose of this work is to investigate the use of several machine learning models, such as Support Vector Machine, Gaussian Naive Bayes, Bernoulli Naive Bayes, Random Forest, and Logistic Regression, for the identification of lung cancer. We compare these models in order to analyze how well they work, identify their advantages and disadvantages, and offer some insight into how they might be used in clinical settings. Our goal is to help improve patient outcomes and lessen the impact of lung cancer on the world's health by tackling the problem of early lung cancer detection.

**Importance of early detection**

1. More Treatment Choices   
2. Higher Rates of Survival   
3. Decreased Mortality and Morbidity   
4. Minimized Side Effects of Treatment   
5. Reduced Medical Expenses   
6. Possibility of Recovery

1. **Research Gap Identified**

There is a large research gap because of the limits of the current methods for lung cancer detection, which include early screening programs and conventional diagnostic tools. Among these restrictions are:

* 1. Low Specificity and Sensitivity
* 2. Radiation exposure and invasive procedures
* 3. High Rates of False Positives
* 4. Heavily Requires Resources
* 5. Inadequate Utilization

**Proposed system**

Our study aims to assess several machine learning models' efficacy in lung cancer early detection. With regard to precisely identifying lung cancer based on features taken from medical imaging scans and patient demographic data, our specific goal is to evaluate the efficacy of Random Forest, Logistic Regression, Gaussian Naive Bayes, Bernoulli Naive Bayes, and Support Vector Machine models.

**Pre-Processing for ML Models**

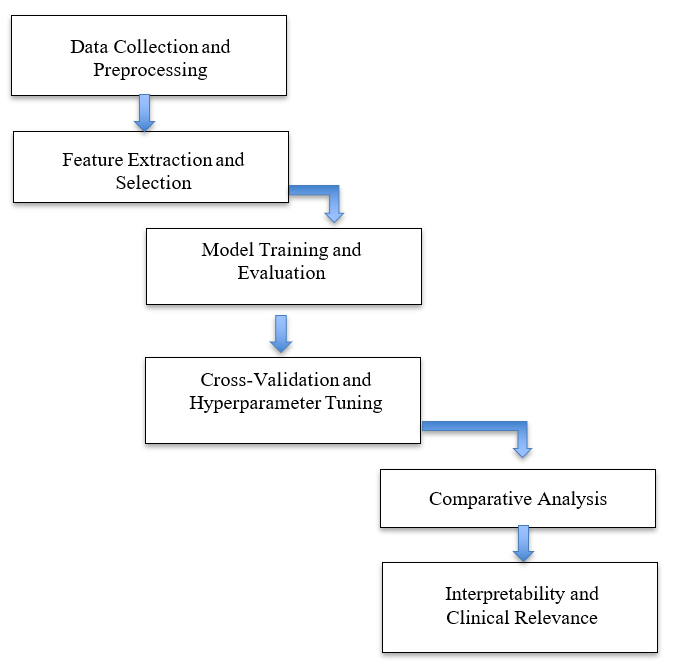
****

Figure1: Flow chart for Pre-processing

1. **Methodology**

The dataset that we employed for our study includes a range of personal health and demographic characteristics, with the goal of investigating potential risk factors for lung cancer. Details like gender (male or female), age in years, smoking status (represented by '1' for non-smokers and '2' for smokers), anxiety levels, peer pressure susceptibility, chronic disease presence, fatigue, allergies, wheezing, alcohol consumption habits, coughing, shortness of breath, difficulty swallowing, and chest pain are all included. 'LUNG\_CANCER', a target variable in the dataset, indicates whether or not the individual has received a lung cancer diagnosis ('1' for no and '2' for yes).

**Algorithms used for analysis**

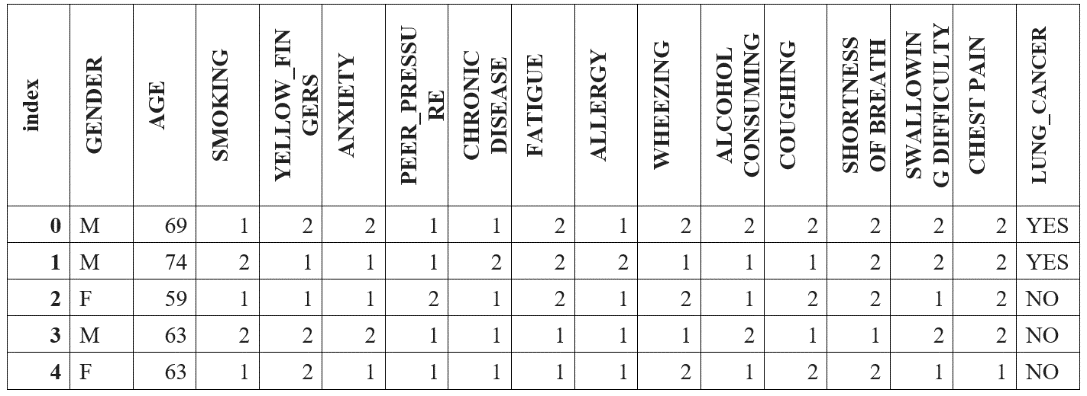
Lung cancer detection using machine learning models involves leveraging algorithms to analyze data and predict the likelihood of lung cancer based on various features:

1. **Random Forest:** This ensemble learning method creates predictions by combining several decision trees. Individual predictions are made by each decision tree once it has been trained on a random portion of the data. The total of all the trees' projections determines the final prediction. In order to categorize people as having or not having lung cancer, Random Forest can find patterns in information such patient demographics and results from medical imaging exams.
2. **Logistic Regression Model:** A statistical model used for binary classification tasks is called logistic regression. It calculates the likelihood of a binary result by taking into account one or more independent variables. Logistic regression can be trained on characteristics like age, smoking status, and medical history to determine a person's risk of developing lung cancer.
3. **Gaussian Naive Bayes Model:** The Naive Bayes algorithm is a probabilistic classifier that relies on the premise of conditional independence of features and the Bayes theorem. Continuous features are assumed to follow a Gaussian distribution in the Gaussian Naive Bayes variant. Gaussian Naive Bayes can be trained on characteristics like age, smoking status, and symptoms in order to determine a person's likelihood of getting lung cancer.
4. **Bernoulli Naive Bayes Model:** Bernoulli Naive Bayes is a probabilistic classifier that is intended for binary features, much as Gaussian Naive Bayes. It makes the assumption that characteristics, like present or absent, are binary variables. Bernoulli Naive Bayes can be used to analyze features that show whether symptoms or risk factors for lung cancer are present or absent.
5. **Support Vector Machine Model:** For problems involving regression and classification, the supervised learning algorithm Support Vector Machine (SVM) is employed. In order for SVM to function, the feature space's optimum hyperplane for class separation must be located. SVM can be trained on features, such as imaging scans and patient data, with the purpose of classifying people into groups according to whether they are positive or negative for lung cancer based on feature representations.
6. **Experimental Setup:**

Lung cancer is well known for its high death rate and aggressive behavior, which frequently results in significant tissue damage and impacts nearby organs. Lung cancer is a major cause of cancer-related death worldwide, and because it is so elusive, it presents a tough challenge. It develops when aberrant cells multiply out of control in the lungs, leading to tumors that have the potential to spread to other parts of the body. Lung cancer progresses slowly, making it difficult to detect early on. As a result, patients are frequently diagnosed at advanced stages, when there are few treatment options and a bad prognosis. The innate challenge of timely identification highlights the pressing requirement for efficient diagnostic techniques and therapeutic approaches to enhance the prognosis of persons susceptible to or impacted by lung cancer.

**Sample Dataset**

**Table 1: Input Dataset**



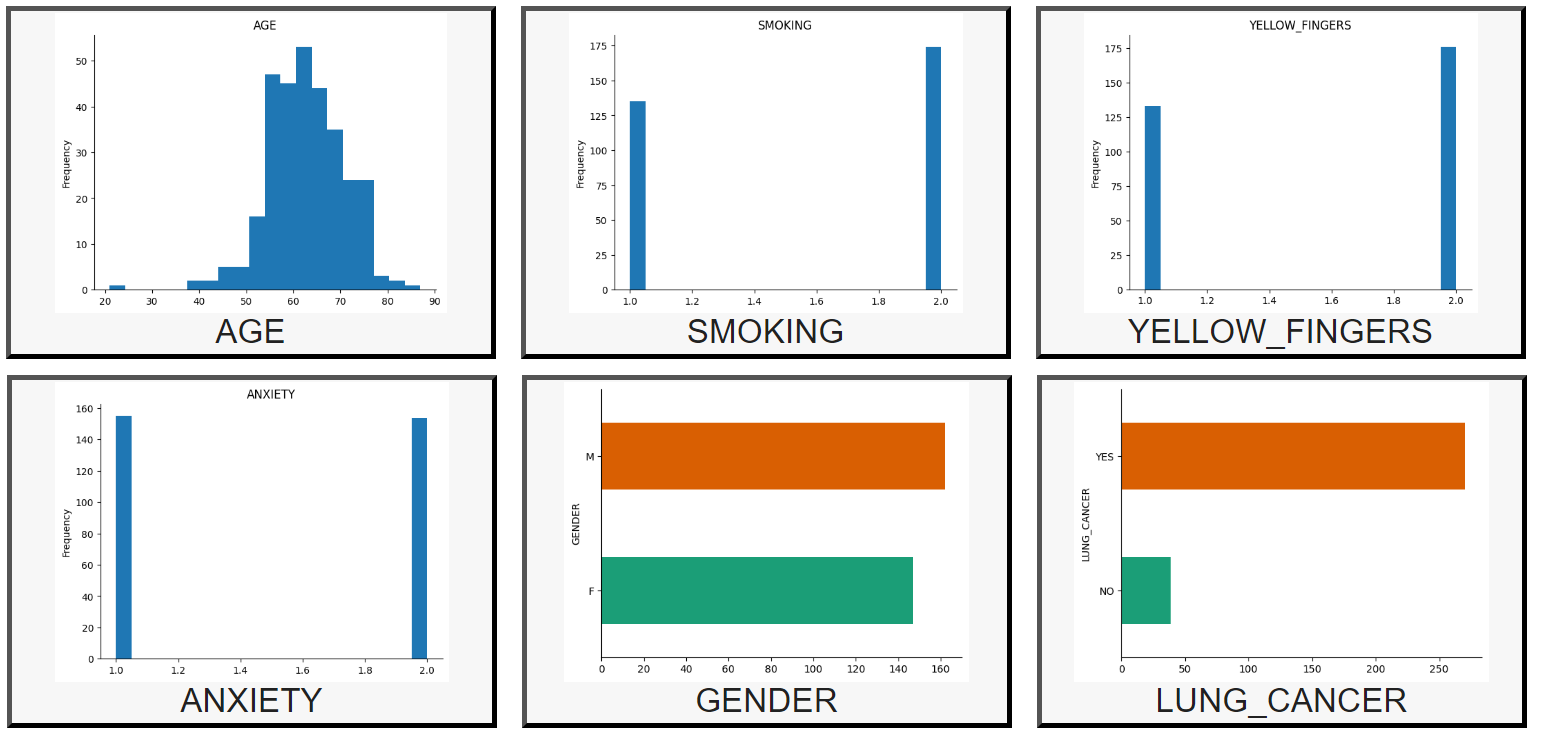


Figure 1: Sample image Dataset

**Visual Representation of Data Distribution: Histogram Analysis**

A histogram is a graph that shows how data values are distributed within predetermined bounds. It displays data as bars, with each bar's height representing the frequency or percentage of data points falling inside pre-established ranges or bins. Histograms help analysts spot patterns, trends, and anomalies by giving insights into the form, central tendency, and variability of a dataset through data visualization. Within the framework of our research, we utilize histograms to examine the distribution of pertinent variables under particular circumstances, enabling a more profound comprehension of the fundamental features of the data and guiding ensuing data-driven judgments and analyses. The "YES" or "NO" parameters were used in the analysis.

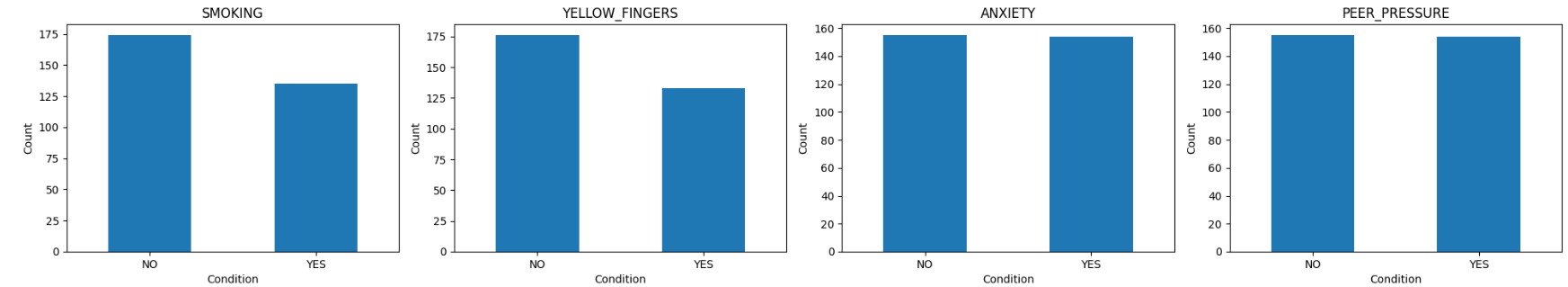


Figure 2: **Histogram Analysis Chart**

**Circular Histogram Analysis for ‘YES’ Frequency for each parameter**

A circular histogram is a visualization technique that shows the distribution of data values around a circular axis. It is sometimes referred to as a rose diagram or circular bar chart. In particular, we pay attention to how frequently 'YES' replies occur for each parameter in our dataset. All the parameters are shown as segments around the circle, and the length of each segment shows the percentage of 'YES' responses that were seen. We can efficiently display the relative prevalence of "YES" replies across various parameters by utilizing a circular histogram, which offers insightful information about the distribution patterns and any correlations within our dataset. With the use of this visualization technique, we can get more in-depth understandings of the 'YES' frequency data by identifying trends, clusters, and outliers.

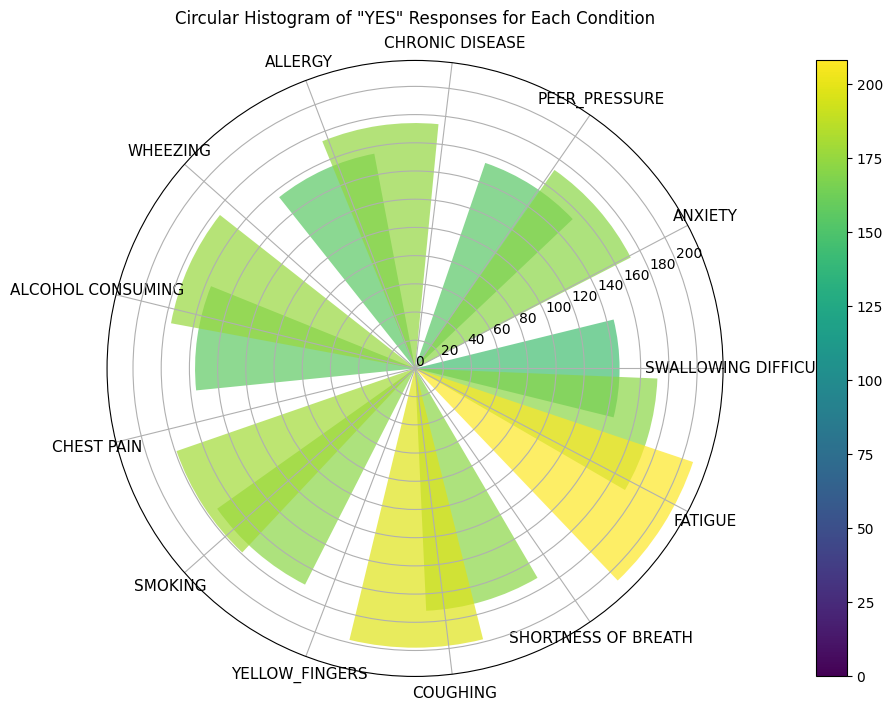


Figure 3: Rose diagram

**Classification analysis of Various algorithms**

**Tabe 2: Table of Comparisons of valid output parameters**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **#** | **Algorithm** | **P** | **R** | **F1** | **S** | **A** |
| **1** | Logistic Regression Model (LR Model) | 0.97 | 0.97 | 0.97 | 62 | **96.77%** |
| **2** | Gaussian Naive Bayes Model (G-NB Model) | **0.94** | **0.95** | **0.94** | **62** | **95.16%** |
| **3** | Bernoulli Naive Bayes Model (B-NB Model) | **0.97** | **0.97** | **0.97** | **62** | **96.77%** |
| **4** | Support Vector Machine Model (SVM Model) | **0.96** | **0.94** | **0.95** | **62** | **93.55%** |
| **5** | Random Forest Model (RF Model) | **0.97** | **0.97** | **0.97** | **62** | **96.77 %** |

*\*. P- Precision, R- Recall, F1 – F1 Score, S- Support, A- Accuracy*

The result comparison says that 3 algorithms LR Model, B-NB Model, RF model provided the results with an accuracy of 96.77% with a moderate F1 Score, Precision, recall & support.

1. **Results:**

We report the following important conclusions from our analysis of the efficacy of different machine learning models for lung cancer detection- With accuracy scores above 96.77%, the Random Forest, Bernoulli Naive Bayes, and Logistic Regression models showed the best resilience and accuracy in identifying people with and without lung cancer. With accuracy ratings ranging from 90% to 95%, the Gaussian Naive Bayes Model and Support Vector Machine Model both demonstrated strong performance. Although the accuracy of the Support Vector Machine Model was marginally lower than that of the other models, it nevertheless performed admirably, especially in situations when the features were binary.

1. **Future Directions:**

It is necessary to do future study to confirm the results in bigger and more varied datasets, taking into account additional clinical and genetic factors that could affect the risk of lung cancer. The study of model ensembling and ensemble learning methodologies may improve the generalizability and robustness of lung cancer detection models. To effectively deploy and adopt machine learning-based techniques for lung cancer screening and diagnosis, collaboration with healthcare institutions and stakeholders is crucial in translating research findings into clinical practice.

1. **Conclusion**

By examining multiple models, such as Random Forest, Logistic Regression, Gaussian Naive Bayes, Bernoulli Naive Bayes, and Support Vector Machine, we have discovered significant results that highlight the potential of machine learning to transform the detection and treatment of lung cancer.

Our research showed that the best models for diagnosing people with lung cancer were the Random Forest, Bernoulli Naive Bayes, and Logistic Regression models, all of which had accuracy rates above 96.5%. Furthermore, our study showed that the most significant predictors in all models were age, the prevalence of chronic conditions, and smoking status. These findings underscore the significance of taking into account a complete collection of features in predictive modeling and offer insightful information about the underlying determinants linked to lung cancer risk.

**Acknowledgement**

We express our gratitude for the great help and encouragement we have received from so many people, which has allowed us to successfully finish this project. Respectful Gratitude to Dayananda Sagar University Main Campus (Harohalli) Department of Computer Science & Engineering (Artificial Intelligence & Machine Learning). Additionally, we appreciate and are appreciative of Kaggle for supplying the necessary dataset for the analysis.

**Bibliography:**

[1] Raoof, Syed Saba, M. Jabbar, Syed Aley Fathima. "Lung Cancer Prediction using Machine Learning: A Comprehensive Approach." 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA). IEEE, 2020.

[2] S. Luxmi, J Kaur Sandhu, and N Goyal. "Intelligent Method for Detection of Coronary Artery Disease with Ensemble Approach." Advances in Communication & Computational Technology, Singapore, 2021 Springer. 1033-1042.

[3] Banerjee, Nikita, and Subhalaxmi Das. "Prediction Lung Cancer–In Machine Learning Perspective." 2020 International Conference on Computer Science, Engineering & Applications (ICCSEA). IEEE, 2020.

[4] Mathur, Prashant, et al. "Cancer statistics, 2020: report from national cancer registry programme, India." JCO Global Oncology6: 1063-1075.

[5] Singh G. et al. "Comparison between multinomial and Bernoulli naïve Bayes for text classification." 2019 International Conference on Automation, Computational and Technology Management (ICACTM). IEEE, 2019

[6] Xing, PuǦYuan, et al. "What are the clinical symptoms and physical signs for nonǦsmall cell lung cancer before diagnosis is made? A nationǦwide multicenter 10Ǧyear retrospective study in China." Cancer medicine 8.8 (2019): 4055-4069

[7] Sung Hyuna, et al. “Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries”. CA: A Cancer Journal for Clinicians: Available online at cacancerjournal.com.

[8] D. Li, Q. Ge, P. Zhang, Y. Xing, Z. Yang and W. Nai, "Ridge Regression with High Order Truncated Gradient Descent Method," 2020 12th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), Hangzhou, China, 2020, pp. 252- 255.

[9] F. Kabir, S. Siddique, M. R. A. Kotwal and M. N. Huda, "Bangla text document categorization using Stochastic Gradient Descent (SGD) classifier," 2015 International Conference on Cognitive Computing and Information Processing(CCIP), 2015, pp. 1-4.

[10] Patra, Radhanath. "Prediction of Lung Cancer Using Machine Learning Classifier." International Conference on Computing Science, Communication and Security. Springer, Singapore, 2020.

[11] M. Uppal, R. Kaur, D. Gupta, and K. Gupta, “Determinants of COVID19 disease severity in patients having diabetes,” AIP Conf. Proc., vol. 2357, no. 1, p. 040011, May 2022, doi: 10.1063/5.0080607.

[12] S. Sharma, K. Guleria, S. Tiwari, and S. Kumar, “A deep learning based convolutional neural network model with VGG16 feature extractor for the detection of Alzheimer Disease using MRI scans,” Measurement: Sensors, vol. 24, p. 100506, Dec. 2022, doi: 10.1016/j.measen.2022.100506.

[13] K. Guleria, S. Sharma, S. Kumar, and S. Tiwari, “Early prediction of hypothyroidism and multiclass classification using predictive machine learning and deep learning,” Measurement: Sensors, vol. 24, p. 100482, Dec. 2022, doi: 10.1016/j.measen.2022.100482.

[14] P. M. Shakeel, M. A. Burhanuddin, and M. I. Desa, “Automatic lung cancer detection from CT image using improved deep neural network and ensemble classifier,” Neural Comput. Appl., vol. 34, no. 12, pp. 9579–9592, Jun. 2022, doi: 10.1007/s00521-020-04842-6.

[15] R. Sujitha and V. Seenivasagam, “RETRACTED ARTICLE: Classification of lung cancer stages with machine learning over big data healthcare framework,” J. Ambient Intell. Humaniz. Comput., vol. 12, no. 5, pp. 5639–5649, May 2021, doi: 10.1007/s12652-020-02071-2.