# Lung Cancer Detection using Machine Learning Techniques

**Megha Rani Raigonda**
Assistant Professor, Department of Computer Science and Engineering (MCA), Visvesvaraya Technological University Kalaburagi, Karnataka
E-mail: megharaigond@rediffmail.com
ORCID: https://orcid.org/0000-0001-9964-3265

**Girish Mama**
Post Graduate Student, Department of Computer Science and Engineering (MCA), Visvesvaraya Technological University Kalaburagi, Karnataka
E-mail: girishmama120@gmail.com
ORCID: https://orcid.org/0009-0008-1988-9520

**Rajkumar Bainoor**
Assistant Professor, Department of Electronics and Communication Engineering, PDA college of Engineering, Kalaburagi, Karnataka
Email: rajmolkera@gmail.com

**Abstract :** Lung cancer (Bronchiogenic Carcinoma) is a serious worldwide health issue, and detecting it early is essential to enhance patient outcomes each year. The problem is that lung cancer is typically found after it has spread, giving patients a poor prognosis and few treatment options. The recommended approach is centered on developing a patient-specific lung cancer screening system. The purpose of the study is to diagnose important lung cancer risk factors and warning signs, such as cigarette smoking history, environmental exposure, occupational hazards, family medical history, and chronic respiratory conditions. A large and diverse dataset thatis obtained from Kaggle and pre-processed using a number of techniques is used to create the suggested model. A few of the ML approaches employed in the suggested methodology are XG Boost, Decision Tree, K Nearest Neighbour, and Random Forest Classifier, the latter of which has the greatest accuracy of 95.16%.

**Index Terms:** Lung Cancer detection, Machine Learning, Kaggle, XG Boost, Decision Tree, KNN, Random Forest.

## 1. Introduction

The fatal illness known as cancer is typically brought on by a number of pathological changes and inherited abnormalities. Multiple organs are affected by cancer at the same time, and diverse cancer types can develop in different bodily organs [1]. Anywhere in the body, cancerous cells have the potential to develop atypically and endanger life. When symptoms first appear, it must be accurately and quickly recognized in order to select the most appropriate course of therapy. Tumor-forming carcinomatous proliferation is currently the leading cause of mortality worldwide. The majority of fatalities are caused by Bronchiogenic Carcinoma (1.6 million) [2] among the several forms of cancer that are currently recognized, including breast, liver, stomach, and colorectal. Of those, 75% die within five years after diagnosis. Despite the fact that every approach has a unique set of problems, a challenging past, and insufficient diagnosis. Additionally, early stage patients who got appropriate therapy were said to have a 40% probability of staying alive for 5 years [3]. When you breathe in, air enters your mouth or nose and moves via your trachea (windpipe) to your lungs. About 10-15% of Bronchiogenic Carcinoma cases are found in persons who have never smoked, and the great majority (85%) of cases are caused by long-term tobacco use [4]. The intricacy of cancer cells, which leads to medication resistance, and the high intra-tumor heterogeneity (ITH) make treating cancer more difficult. Smaller bronchi from the trachea divide into the lungs before separating once again. Continuous technological advancement over the past few decades has made it possible to create a large number of clinical, medical imaging, and genetic databases as well as a large number of substantial collaborative cancer endeavors. Your lungs' primary functions are to absorb oxygen and exhale carbon dioxide. Chemotherapy is the choice with the greatest long-term results for NSCLC patients with the podium IIIB-IVB illness progression and efficacy circumstance 2 [5]. Compared to SCLC, NSCLC is more prevalent and tends to progress and spread more slowly [6]. The cells that line the bronchi and other areas of the lung, such as the bronchioles or alveoli, are typically where Bronchiogenic Carcinoma first develops. During breathing, the diaphragm extends and contracts, forcing air into and out of the lungs.

A healthy pulmonary system are both tactile and visual like absorbers. It is believed that their exteriors are brightly colored, squishy, & adaptable could encourage them to puff up and exceed with each exhalation & inhalation. Their main duty is to circulate fresh air throughout your arterial system via the air you airflow. When you take a breathe, air circulates in into your system via a passageway that link the nostrils, throat, and the pulmonary system with your bronchus. Bronchial passageways are subsequently utilized to funnel the air, paving the way it to seek employment in & escape the pulmonary system. Along your airways, sputum and tiny structures resembling hairs indicated follicles abundantly obvious debris alongside various matter that occur in via the breeze. Effective since it attains the air passageways, or miniatures, balloon-like balloons in the pulmonary system, breeze follows up to navigate onto the passageways. Much of the oxygen then penetrates the circulatory via that area. At that point respire, the pulmonary system destroy gaseous form from your pulse via an endeavor deemed gas movement. Consuming cigarettes hurls off their equilibrium of the full workflow.

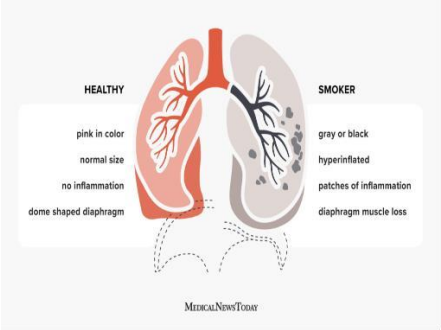



Fig.1. Difference between Healthy and Cancerous Lung

Nearly 70 of the more than 7,000 compounds included in a single cigarette puff are recognized carcinogens. These poisons enter your lungs through your breath and irritate them. Too much mucus starts to build up in your airways. That results in issues including pneumonia, bronchitis, and coughing. Your lungs' small airways enlarge as a result of toxins. This can cause wheezing and shortness of breath as well as a tightening sensation in your chest. Smoking can cause inflammation to turn into scar tissue, which makes it more difficult to breathe. You also develop a buildup of tobacco's sticky tar within your lungs. After several years of smoking, they may become black. Cigarette smoke contains nicotine, which temporarily paralyzes and destroys cilia. This implies that the dust and filth in the air you breathe cannot be filtered by your airways. Additionally, it increases your risk of developing colds and other respiratory illnesses. Smoking also harms your body's alveoli, which are small air sacs that carry oxygen throughout it. After being destroyed, they stop growing. Emphysema, a lung disorder that causes severe breathlessness, develops when you lose too many of them. Smoking exposes all of your integrity of your critical tissues in imminent danger by diminished the sheer amount of oxygen it gathers andbettering the extent of the poisonous gas it encounters.

The size of the original tumor, how deeply it penetrates the surrounding tissue, and whether it has migrated to the lymph nodes or other organs are the main factors used to stage cancer. There are specific staging recommendations for each form of cancer.

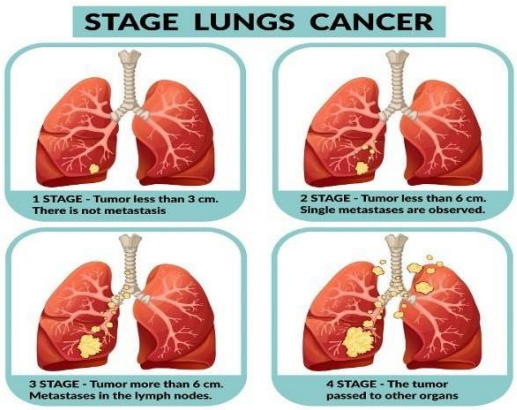



Fig.2. Stages of Lung Cancer

There are a number of possible size and spread combinations at each level. For instance, a Stage III cancer may have a smaller initial tumor than a Stage II cancer, but additional variables may have progressed the cancer to a more serious stage. The overall lung cancer staging is as follows:

- Stage 0 (in-situ) : Carcinoma is found in the foremost layer of the bronchus in stage 0. It hasn't migrated to the exterior or to other bronchus tissue.
- Stage 1 : Stage 1 tumors are limited to the lung. It is somewhat little (less than 4 centimeters). It hasn't migrated outside the chest or to surrounding lymph nodes. Surgery is generally used to treat stage 1 malignancies. For patients who are not candidates for surgery, we provide radiation treatment. Chemotherapy, targeted treatment, and immunotherapy are typically not necessary for stage 1 patients.
- Stage 2 : Larger tumors (greater than 4 cm in diameter) are present in stage 2 lung cancer. Or there are indications that the cancer has not migrated outside the lung but rather to surrounding lymph nodes. Stage 2 lung cancer is often treated surgically, then with chemotherapy, targeted treatments, or immune therapies. The individuals who are forbidden from amputation are frequently given chemotherapy and radiation as a substitute.
- Stage 3 : Stage 3 Bronchiogenic Carcoinoma is characterized by the presence of malignancy in the chest lymph nodes farthest from the lung. Large tumors that have migrated to neighboring lymph nodes can also exist. Most persons with stage 3 cancer receive many types of care. This may involve a mix of immunological therapies, targeted therapies, surgery, radiation, and chemotherapy.
- Stage 4 : Stage 4 Bronchiogenic Carcinoma spreads beyond the thoracic cavity, where it first appeared. The latter includes the pulmonary, the ribs, the hippocampus, and the gallbladder, which is an organ above the urinary system are the most often affected organs. The kind of tumor will determine the course of treatment. Chemotherapy, immunological or targeted treatments, or a combination of them, may be used.

ML, a branch of AI, looks to make predictions by utilizing mathematical algorithms to find patterns in data. The suggested approach focuses on early lung cancer detection based on a person's health and behaviors, which improves patient outcomes and increases the possibility that a successful course of therapy would be followed. The suggested method for identifying lung cancer takes into account the health condition and routine habits of patient. The proposed model applies XG Boost, DT, KNN, and RFC on the Kaggle pre-processed dataset.

A lung cancer detection project's problem statement often outlines the specific issue or difficulty that the initiative seeks to address. A project on lung cancer detections' problem statement is: Bronchiogenic Carcinoma carries on being the biggest cause of Carcinoma-related fatalities globally, despite improvements in imaging and diagnostic technologies. The present issue is that lung cancer is frequently discovered late, leaving patients with a poor prognosis and few treatment options. Additionally, the broad use of early detection techniques is hampered by the absence of specialized procedures and integration into healthcare systems. Therefore, it is essential to provide cutting-edge, precise, and trustworthy lung cancer detection technologies that allow for early detection, individualized therapy, and improved patient survival rates.

Depending on its particular objectives and scope, a lung cancer detection project's goals may alter. However, the following are some common objectives that a project of this nature could attempt to achieve: The development of techniques or technologies that can identify the disease early is among the primary goals of Bronchiogenic Carcinoma detection research. Prompt recognition optimizes results of patients & increases the likelihood that a medication will be effective. Biomarker analysis and diagnostic visual inspection (including radiation therapy and the cerebral cortex examinations) are used to increase accuracy and reliability. The primary goal is to lower false positives and inaccurate outcomes. to create cutting-edge methods or tools made particularly for detecting lung cancer. to aid in the understanding of medical images. To improve the suggested model's accuracy in order to generate accurate results and reduce false positives and false negatives.

Lung cancer can currently be diagnosed using the patient's regular blood markers. Blood tests may be a useful tool during the entire evaluation process, despite the fact that they are not consistently able to identify lung cancer on their own. The complete blood count (CBC), tumor markers, inflammatory markers, tests for liver and kidney function, among other blood indicators, are all taken into account by the current lung cancer screening.

technique. Despite the fact that lung cancer may be detected by blood tests. The hazards and limitations of relying only on lung cancer screening are numerous.

The following are some negatives: Lack of specificity, falsely negative results, a restricted diagnostic utility, interference from other factors, a low predictive value, cost, and accessibility are a few of the problems.

The system's main objective is early diagnosis of lung cancer based on a individual's lifestyle and health. It enhances patient outcomes and raises the likelihood that a therapy will be effective. The recommended strategy for detecting lung cancer takes into account the health condition and routine habits of patient. The recommended model applies many methods, such as XG Boost, DT, KNN, and RFC to the pre-processed dataset that is received from Kaggle. The suggested method may be used for various things, such as better accuracy, risk assessment, early detection, tailored approach, quick triage, and more.

## 2. Related Work

[1] Dakhaz Mustafa Abdullah and others investigated the efficacy of classification algorithms for determining the severity of the lung cancer illness using the WEKA Tool. [2] B R Manju suggested a model to look at the importance of the pre-cancerous stage in the early diagnosis of lung cancer. [3] A novel multidisciplinary approach was put forth by Ying Xie and Wei-Yu Meng used metabolomics and machine learning techniques to find early lung cancer detection indicators. [4] Negar Maleki and others employed kNN, a machine learning technique, in combination with a feature-selection genetic algorithm to divide lung cancer risk into three categories: low, medium, and high. [5] Hunter A. Miller and Xinmin Yin laid the groundwork for comprehensively evaluating the metabolic properties of NSCLC tissue generated from patients with the longer-term goal of predicting the response of individual patients to first-line treatment regimens. [6] V.Krishnaiah and othes created a prototype Bronchiogenic Carcinoma illness system to identify the target population that requires further screening for lung cancer disease, in order to lower the prevalence and fatality rate. [7] Quoc-Nam Tran proposed a strategy that may locate reasonably priced biological indicators as quantitative metrics for nearly flawless NSCLC lung cancer prediction accuracy. [8] Qing Wu and Wenbing Zhao suggested a vectorized histogram feature-based EDM machine learning approach to detect SCLC for early cancer prediction. [9] Shubhada Agarwal, Sanjeev Thakur, Alka Chaudhary performed research at COLAB utilizing a lung cancer dataset to develop a machine learning model with 13 parameters. [10] Mohseena Thaseen, S.K.UmaMaheswaran and Darshana A Naik Since it can analyze cancer tissue from histopathological pictures and CT scan images, CNN has been proposed as a tool for cancer diagnosis. [11] Abhishek Verma, Cabinet Kumar Shah and Veerpal Kaur Supervised Learning Techniques, which are extensively used classification algorithms, are utilized to forecast all cancer kinds. [12] Ganta Sruthi, Chokkakula Likitha Ram and Malegam Koushik Sai created a cooperative plan for identifying thisillness and communicating the patient's situation. [13] Amit Singh, Rakesh Kumar and Rajul Rastogi used a number of machine learning algorithms, including the Ensemble classifier, Multinomial RFC, SVM, KNN, NB, and SGD. [14] Harikumar Rajaguru and others designed CAD system using the Gaussian Mixture Model methodas a classification model was explored. [15] Muntasir Mamun, Afia Farjana XGBoost, LightGBM, AdaBoost, and bagging are four types of ensemble learning approaches that were created to predict lung canc Bronchiogenic Carcinoma using the lung cancer dataset. [16] M.Siddardha Kumar, K.Venkata Rao The way the diagnostic process was created has a major impact. [17] Puneet, Anamika Chauhan developed model that might aid medical professionals in spotting lung cancer and provide confirmation through diagnostic testing, thus saving a person's precious life. [18] S Tandungan said that the LIDC-IDRI dataset was used to train and test the ELM algorithm for identifying lung cancer. [19] D U Wutsqa, A Farhan showed that the testing data somewhat declined and the training data performed well. For an RBFNN model to identify lung cancer with high accuracy, SOM learning is superior to K-means learning. [20] Rana Dhia, Abdu-Aljabar and Osama A XGBoost classification algorithms were used to create a prediction model, which enhanced the accuracy of Bronchiogenic Carcinoma diagnostic detection and relapse prediction. [21] Hamdalla F, Al-Yasri created a reliable computer-aided diagnosis method for quickly identifying this deadly malignancy. [22] Yessi Jusman1, Zul Indra and Roni Salambue demonstrated the effectiveness of a neural network using MLP and RBF architectures for lung cancer data.

## 3. Proposed Methodology

The accuracy and precision of the proposed model are evaluated against those of XG Boost, DT, KNN, and RFC, among other ML techniques.

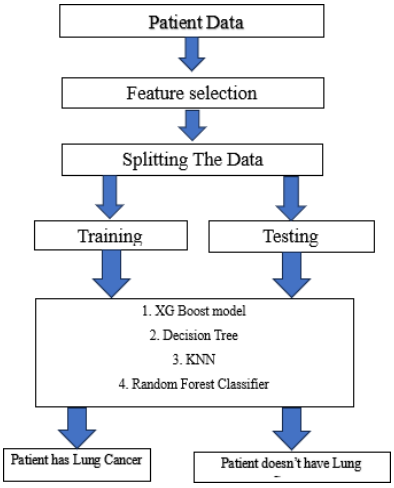
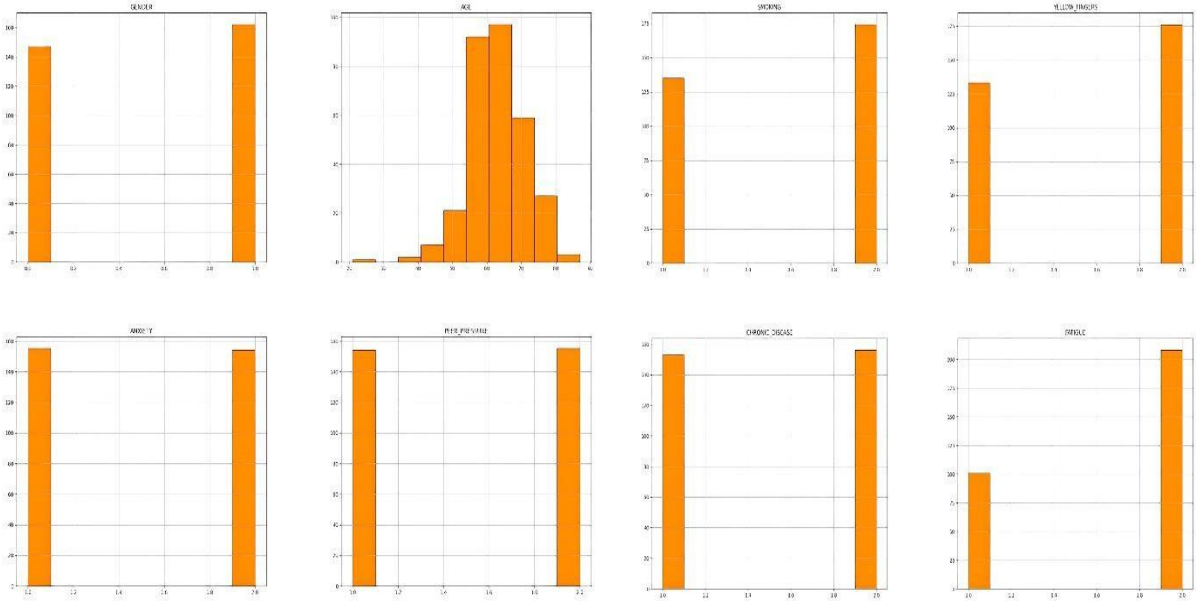



Fig.3. Proposed System Architecture

The main output of this study is a Random Forest Classifier model that outperforms all other Random Forest Classifier models on the pre-processed Dataset in terms of accuracy (95.16%). Here, we talk about the Random Forest Classifier model we developed for detecting lung cancer based on a person's lifestyle and health. The suggested methodology enhances patient outcomes and raises the likelihood of a good outcome. The model may be used for various things, such as higher accuracy, risk assessment, early detection, tailored approaches, quick triage, and more.

## Dataset

Thanks to the pre-processed datasets available on Kaggle, the largest data science community in the world, everyone may fulfill their data science goals without spending any money. Scientific investigations employ a wide range of pre-processed datasets that Kaggle offers.



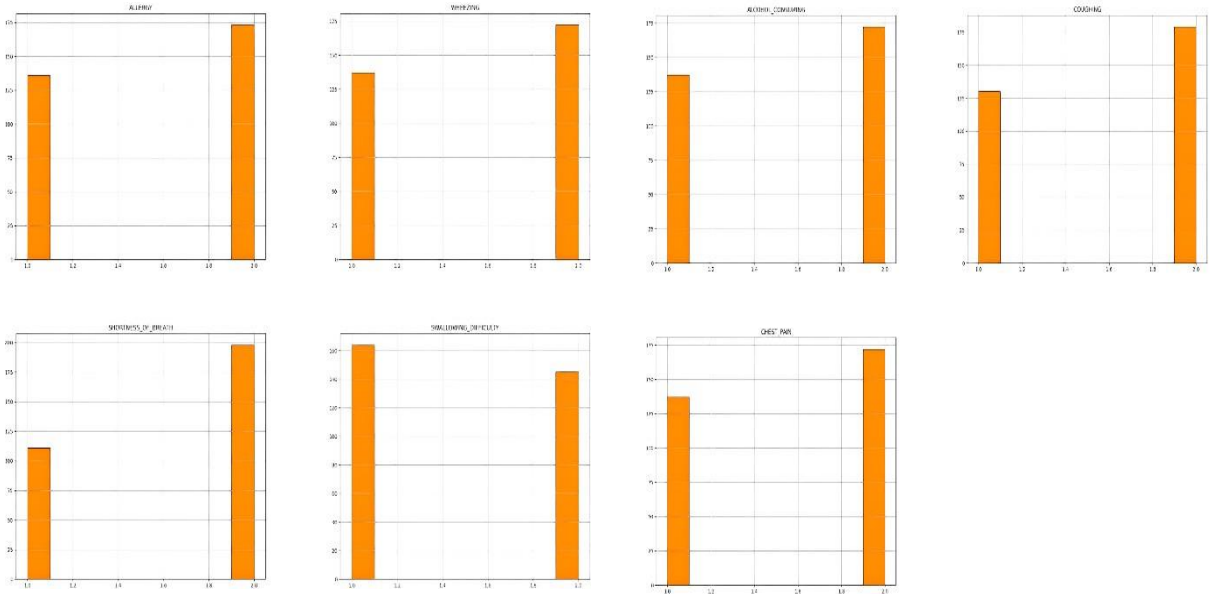



Fig.4. Dataset containing the Lung cancer patient details

The suggested model is developed using a total of about 300 records, 16 factors, including the patient's behaviors and health, are included in the dataset used to create the suggested model. 80% of the data in this model is utilized for training, while the remaining 20% is used for testing.

## 4. Results and Discussion

The number of algorithms used for Bronchiogenic Carcinoma detection is listed in the table below. The table includes a comparison of the various algorithms' implemented percentages for accuracy, precision, recall, and F1 score. Here, it is evident that, when correlated to all other algorithms used, the RFC method obtains the greatest accuracy, precision, recall, and F1 score in the identification of Lung Cancer.

Table 1. Comparison of various Algorithms implemented.

| Algorithm | Accuracy | Precision | Recall | F1 score |
|---|---|---|---|---|
| XG Boost | 91.93% | 75% | 66.6% | 70.5% |
| Decision Tree | 87.09% | 62.5% | 50% | 55.5% |
| KNN | 93.54% | 62.5% | 83.3% | 71.4% |
| Random Forest | 95.16% | 75% | 85.7% | 80% |

The graph below shows unequivocally that the Random Forest algorithm has the best accuracy, at 95.16 percent. The suggested model examined the accuracy of many algorithms, including XG Boost, Decision Tree, K Nearest Neighbors, and Random Forest Classifier, of which the latter obtains the maximum accuracy for diagnosing Bronchiogenic Carcinoma sickness.

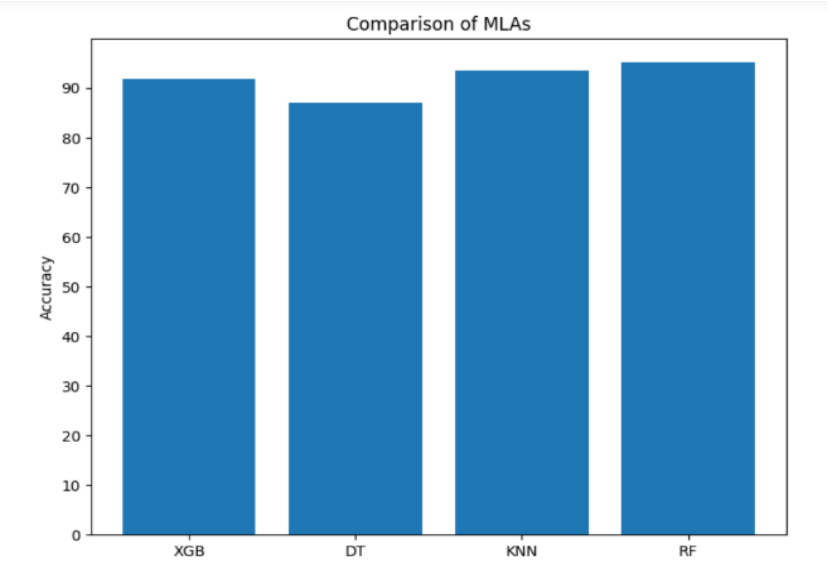



Fig.5. Comparison of Accuracy of various Machine Learning Algorithms implemented

## 5. Conclusion

A number of pathological alterations and hereditary disorders frequently combine to cause the deadly condition known as cancer. This approach's primary objective is to detect Bronchiogenic Carcinoma at an early stage based on a person's lifestyle and overall health. By considering the numerous health and lifestyle factors that influence the development of Bronchiogenic Carcinoma, healthcare professionals can enhance the accuracy of detection & perhaps save lives. The proposed approach, which provides more accuracy than the preceding domain, is built on the Random Forest Classification methodology. The recommended method has a lung cancer detection accuracy of 95.16%.

## References

[1] Dakhaz Mustafa Abdullah, Adnan Mohsin Abdulazeez, Amira Bibo "Lung cancer Prediction and Classification based on Correlation Selection method Using Machine Learning Techniques" Doi: 10.48161/Issn.2709-8206
[2] B R Manju "Efficient multi-level lung cancer prediction model using support vector machine classifier" Sci. Eng. 1012 012034
[3] Ying Xie, Wei-Yu Meng. "Early lung cancer diagnostic biomarker discovery by machine learning methods" Oncology 14 (2021) 100907
[4] Negar Maleki, Yasser Zeinali, Seyed Taghi Akhavan Niaki "A k-NN method for lung cancer prognosis with the use of a genetic algorithm for feature selection" Volume 164, February 2021, 113981
[5] Hunter A. Miller, Xinmin Yin "Evaluation of disease staging and chemotherapeutic response in non-small cell lung cancer from patient tumor-derived metabolomic data" Lung Cancer 156 (2021) 20–30
[6] V.Krishnaiah, Dr.G.Narsimha, Dr.N.Subhash Chandra "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques" IJCSIT, Vol. 4 (1) , 2013, 39 – 45
[7] Quoc-Nam Tran "A novel method for finding non-small cell lung cancer diagnosis biomarkers" BIOCOMP'11, Las Vegas, NV, USA. 18-21 July 2011
[8] Qing Wu and Wenbing Zhao "Small-Cell Lung Cancer Detection Using a Supervised Machine Learning Algorithm" International Symposium on Computer Science and Intelligent Controls, 2017
[9] Shubhada Agarwal , Sanjeev Thakur , Alka Chaudhary "Prediction of Lung Cancer Using Machine Learning Techniques and their Comparative Analysis" 2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO) Amity University, Noida, India. Oct 13-14, 2022
[10] Mohseena Thaseen, S.K.UmaMaheswaran, Darshana A Naik "A Review of Using CNN Approach for Lung Cancer Detection Through Machine Learning" 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)
[11] Abhishek Verma, d Cabinet Kumar Shah, Veerpal Kaur "Cancer Detection and Analysis Using Machine Learning" 2022 Second International Conference on Computer Science, Engineering and Applications (ICCSEA) | 978-1-6654-5834-4/22/$31.00 ©2022 IEEE

[12] Ganta Sruthi, Chokkakula Likitha Ram, Malegam Koushik Sai "Cancer Prediction using Machine Learning" 2022 2nd International Conference on Innovative Practices in Technology and Management (ICIPTM)
[13] Amit Singh, Rakesh Kumar, Rajul Rastogi "Study of Machine Learning Models for the Prediction and Detection of Lungs Cancer" 11th International Conference on System Modeling & Advancement in Research Trends 2022, IEEE.
[14] Harikumar Rajaguru, Sannasi Chakravarthy S R, Sundaresan Chidambaram "Gaussian Mixture Model based Hybrid Machine Learning for Lung Cancer Classification using Symptoms" STCR, 10 – 11 December 2022
[15] Muntasir Mamun, Afia Farjana "Lung cancer prediction model using ensemble learning techniques and a systematic review analysis" DOI: 10.1109/AIIOT54504.2022.9817326
[16] M.Siddardha Kumar, K.Venkata Rao "PREDICTION OF LUNG CANCER USING MACHINE LEARNING TECHNIQUE: A SURVEY" *ICCCI* -2021, Jan. 27 – 29, 2021, Coimbatore, INDIA
[17] Puneet, Anamika Chauhan "Detection of Lung Cancer using Machine Learning Techniques Based on Routine Blood Indices" INOCON, Bengaluru, India. Nov 6-8, 2020
[18] S Tandungan "Comparison of Accuracy in Extreme Learning Machine Based on Hidden Node Structure Variation for Lung Cancer Classification" et al 2019 IOP Conf. Ser.: Mater. Sci. Eng. 676 012014
[19] D U Wutsqa, A Farhan "Lung cancer detection using the SOM-GRR based radial basis function neural network" J. Phys.: Conf. Ser. 1581 012007
[20] Rana Dhia, Abdu-Aljabar Osama A "A Comparative analysis study of lung cancer detection and relapse prediction using XGBoost classifier" OP Conf. Ser.: Mater. Sci. Eng. 1076 012048
[21] Hamdalla F, Al-Yasriy "Diagnosis of Lung Cancer Based on CT Scans Using CNN" IOP Conf. Ser.: Mater. Sci. Eng. 928 022035
[22] Yessi Jusman1, Zul Indra, Roni Salambue "Comparison of Multi Layered Percepton and Radial Basis Function Classification Performance of Lung Cancer Data" et al 2020 J. Phys.: Conf. Ser. 1471 012043

**Authors' Profiles**

Dr. Megha Rani Raigonda, is working as Assistant Professor at the Department of Computer Science and Engineering (MCA), Visvesvaraya Technological University, Centre of PG Studies, Kalaburagi, Karnataka, India. She is in teaching for more than 10 years. She has published more than 30 papers in National / International Journals and presented various papers in National / International conferences. She is author of textbook "Fundamentals of DBMS", Her main area of interest includes Machine Learning, Image Processing, and Artificial Intelligence.

Mr. Girish Mama, a Post Graduation student studying in Department of Computer Science and Engineering (MCA), Visvesvaraya Technological University, Centre for Post Graduation studies, Kalaburagi, Karnataka, India. The main area of interest are ML, DL and AI.

Mr. Rajkumar. P. Bainoor is Assistant Professor in the Dept. of E & CE, P D A College of Engineering, Kalaburagi, Karnataka, India. He is in teaching for more than 15 years. He has published more than 10 papers in International Conferences and journals. His main areas of interest include Power Electronics, PLC, Image Processing,Artificial Intelligence, Computer Networks, and communication System.