IJARSCT

International Journal of Advanced Research in Science, Communication and Technology

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

ISSN: 2581-9429

Volume 6, Issue 3, July 2026

Impact Factor: 8.2

Lung Cancer and its Severity Prediction using Machine Learning

Suchithra V and Dr. Kruthi R

Postgraduate Student, Department of MCA

Associate Professor, Department of MCA

Maharaja Institute of Technology, Mysore

Abstract: *Lung cancer has one of the highest cancer related mortality, therefore early detection is crucial in order to improve survival rates. Machine Learning (ML) algorithms appeared to be a promising solution for predicting lung cancer, by characterizing patients features associated with lifestyle and clinical factors. This paper presents a web-based application for lung cancer and its severity prediction through Machine Learning. A health decision support system was implemented using Flask framework with a MySQL database being used in the background for better efficiency. This study used widely recognised performance metrics to assess and compare the machine learning algorithms implemented in order to identify the most appropriate lung cancer predictive model. Based on the comparative analysis, the better-performing algorithm was selected and integrated into the proposed web application for real-time prediction and severity classification. The system also provides secure user authentication, patient record management, and report generation, making it suitable for practical healthcare applications. The proposed application demonstrates the effective integration of machine learning techniques with web technologies to support timely and reliable preliminary lung cancer prediction..*

Keywords: *Lung cancer*.

1. INTRODUCTION

Lung cancer is one of the most common and life-threatening diseases caused by the uncontrolled growth of abnormal cells in the lung tissues. It remains one of the leading causes of cancer-related mortality worldwide and continues to pose a significant challenge to healthcare systems. Lung cancer is generally classified into two major categories: Non-Small Cell Lung Cancer (NSCLC) and Small Cell Lung Cancer (SCLC), with NSCLC accounting for the majority of diagnosed cases. Early detection of lung cancer is essential because timely diagnosis significantly increases the chances of successful treatment and improves patient survival rates.

Conventional diagnostic methods such as Computed Tomography (CT) scans, Chest X-rays (CXR), Magnetic Resonance Imaging (MRI), sputum cytology, and biopsy are widely used for diagnosing lung cancer. Although these techniques provide reliable diagnostic information, they require specialized medical equipment, experienced healthcare professionals, and considerable processing time.

In this study, two supervised machine learning algorithms, namely K-Nearest Neighbors (KNN) and Gaussian Naïve Bayes (GNB), are implemented for lung cancer prediction and severity prediction using text-based patient datasets. Both algorithms are trained using the same datasets, and their performance is evaluated using accuracy, training time, and prediction time. The experimental results indicate that the KNN algorithm achieved an accuracy of 98.43% for lung cancer prediction, while the Gaussian Naïve Bayes algorithm achieved 97.00%. Similarly, for lung cancer severity prediction, KNN achieved an accuracy of 100.00%, whereas Gaussian Naïve Bayes achieved 91.00%. Based on the comparative performance analysis, the KNN algorithm demonstrated superior prediction accuracy and was therefore selected for deployment in the proposed web application, while Gaussian Naïve Bayes was retained for performance evaluation and comparison.

The proposed system is implemented as a web-based healthcare application using the Flask framework with a MySQL database for secure patient data management. The application provides functionalities including user authentication, patient record management, lung cancer prediction, severity prediction, machine learning model performance comparison, report generation, and discussion forum support. By integrating machine learning techniques with modern web technologies, the proposed system offers a practical healthcare decision-support platform capable of assisting healthcare professionals in performing preliminary lung cancer assessment more efficiently.

II. RELATED WORK

Numerous studies have explored the application of machine learning and deep learning techniques for the early detection and classification of lung cancer. These approaches aim to improve diagnostic accuracy, reduce prediction time, and assist healthcare professionals in identifying lung cancer at an early stage. Various algorithms have been evaluated using both image-based and clinical datasets, demonstrating the growing importance of artificial intelligence in medical diagnosis.

The authors in [5] proposed an automated lung cancer detection system using CT scan images combined with image processing techniques and Support Vector Machine (SVM) classification. Their methodology included image preprocessing, lung segmentation, feature extraction, and classification to distinguish normal and abnormal CT images. The study demonstrated that image processing combined with machine learning can improve the accuracy of lung cancer detection while reducing manual intervention.

In another study, the authors in [6] developed a computer-aided diagnosis system using CT scan images and Convolutional Neural Networks (CNN). The proposed system focused on detecting lung tumors, estimating tumor size, and predicting cancer stages. The authors reported that deep learning techniques significantly improved the accuracy of lung cancer detection while assisting radiologists in clinical diagnosis.

The authors in [7] utilized various supervised learning The authors in [7] presented a deep learning framework for early lung cancer detection using CT images. The proposed methodology incorporated image preprocessing, segmentation, feature extraction, and Convolutional Neural Networks (CNN) to improve tumor identification. The study highlighted that deep learning techniques provide reliable and accurate prediction while reducing diagnosis time and supporting early-stage cancer detection.

The work presented in [8] introduced a DenseNet-based deep learning model optimized using a Genetic Algorithm for lung cancer severity classification. The proposed approach utilized CT scan images to classify benign and malignant lung lesions while optimizing network parameters for improved prediction performance. The study demonstrated that combining deep learning with optimization techniques can enhance classification accuracy and support clinical decision-making.

The authors in [9] proposed a CT image-based lung cancer detection system using Gray Level Co-occurrence Matrix (GLCM) feature extraction and Support Vector Machine (SVM) classification. The system performed image preprocessing, segmentation, feature extraction, and classification to distinguish benign and malignant lung cancer. The study achieved promising classification performance and emphasized the importance of image-based feature extraction in improving diagnostic accuracy.

In another study, the authors in [10] evaluated several machine learning algorithms for lung cancer prediction using text-based clinical datasets. Their research compared different classifiers and employed feature selection techniques to improve prediction performance. The study concluded that selecting an appropriate machine learning model and optimizing feature selection significantly enhanced classification accuracy for lung cancer prediction and severity analysis.

The authors in [11] proposed an Optimized Weighting-Based Enhanced Neural Network (OWENN) model for lung cancer prediction using CT images. Their methodology integrated adaptive median filtering for image preprocessing, optimization techniques for feature extraction, and neural network-based classification. Experimental results demonstrated improved prediction accuracy and reliable classification of cancerous and non-cancerous cases.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, July 2026
Impact Factor: 8.2

The study presented in [12] compared multiple machine learning algorithms, including Random Forest, Decision Tree, Logistic Regression, and Support Vector Machine, for lung cancer prediction. The models were evaluated using standard performance metrics such as accuracy, precision, recall, and harmonic mean. The comparative analysis highlighted the importance of selecting an appropriate classifier to improve diagnostic performance in lung cancer prediction systems.

The authors in [13] proposed a real-time lung cancer detection framework that combines Generative Adversarial Networks (GAN) with Convolutional Neural Networks (CNN) for processing Low-Dose Computed Tomography (LDCT) images. The GAN model was used to enhance image quality, while the CNN model performed lung cancer classification. The study reported improved sensitivity, specificity, and overall diagnostic accuracy for real-time clinical applications.

Although these studies have demonstrated promising results using image processing, machine learning, and deep learning techniques, many of them primarily focus on CT image analysis or evaluating prediction algorithms independently. Comparatively fewer studies integrate machine learning algorithms into a complete web-based healthcare application that supports patient management, lung cancer prediction, severity prediction, algorithm performance comparison, and report generation. To address this limitation, the proposed work develops a Flask-based web application integrated with a MySQL database, implementing both K-Nearest Neighbors (KNN) and Gaussian Naïve Bayes (GNB) for lung cancer and severity prediction. Based on comparative evaluation, the KNN model achieved higher prediction accuracy than Gaussian Naïve Bayes and was therefore selected as the final deployed model in the proposed healthcare decision-support system.

III. METHODOLOGY

A. Data Collection

The proposed methodology focuses on developing a web-based machine learning application for predicting lung cancer and its severity using clinical and symptom-based datasets. The system consists of five major phases: data collection, data preprocessing, data splitting, machine learning model implementation, and web application development. Initially, the datasets are collected and prepared for analysis through preprocessing techniques. The processed data is then divided into training and testing sets to build and evaluate the machine learning models. Finally, the trained model is integrated into a Flask-based web application that enables lung cancer prediction, severity prediction, patient management, and report generation. The overall workflow of the proposed methodology is illustrated in Fig. 1.

TABLE I FEATURES DESCRIPTION OF SURVEY LUNG CANCER DATASET

Features	Description	Data Type
Age	Indicates the age of the patient.	Numeric
Gender	Indicates the gender of the patient (Male/Female).	Nominal
Blood Pressure	Indicates the blood pressure level of the patient.	Numeric
Weight	Indicates the body weight of the patient.	Numeric
Sugar Tested	Indicates whether the patient's blood sugar test result level.	Nominal
Family History	Indicates whether the patient has a family history of lung cancer.	Nominal
Smokes	Indicates whether the patient is a smoker.	Nominal
Alcohol	Indicates whether the patient consumes alcohol.	Nominal
Anemia	Indicates whether the patient has anemia.	Nominal
Heart Disease	Indicates whether the patient has heart disease.	Nominal
Diabetes	Indicates whether the patient has diabetes.	Nominal
Cholesterol	Indicates the cholesterol level of the patient.	Numeric
Result	Indicates whether the patient is diagnosed with lung cancer (Yes/No).	Nominal

Features Description Data Type

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, July 2026
Impact Factor: 8.2

TABLE II FEATURES DESCRIPTION OF CANCER PATIENTS DATASET

Features	Description	Data Type
Index	Unique record number of each patient.	Numeric
Patient Id	Unique identification number of the patient.	Nominal
Age	Indicates the age of the patient.	Numeric
Gender	Indicates the gender of the patient (Male/Female).	Nominal
Air Pollution	Indicates the level of air pollution exposure.	Numeric
Alcohol Use	Indicates the level of alcohol consumption.	Numeric
Dust Allergy	Indicates the severity of dust allergy.	Numeric
Occupational Hazards	Indicates exposure to occupational hazards.	Numeric
Genetic Risk	Indicates the genetic risk of lung cancer.	Numeric
Chronic Lung Disease	Indicates the presence of chronic lung disease.	Numeric
Balanced Diet	Indicates the quality of the patient's diet.	Numeric
Obesity	Indicates the obesity level of the patient.	Numeric
Smoking	Indicates the patient's smoking level.	Numeric
Passive Smoker	Indicates exposure to passive smoking.	Numeric
Chest Pain	Indicates the severity of chest pain.	Numeric
Coughing of Blood	Indicates the severity of coughing blood.	Numeric
Fatigue	Indicates the level of fatigue experienced by the patient.	Numeric
Weight Loss	Indicates the degree of weight loss.	Numeric
Shortness of Breath	Indicates the severity of breathing difficulty.	Numeric
Wheezing	Indicates the presence of wheezing symptoms.	Numeric
Swallowing Difficulty	Indicates difficulty in swallowing.	Numeric
Clubbing of Finger Nails	Indicates the presence of finger clubbing.	Numeric
Frequent Cold	Indicates the frequency of common cold symptoms.	Numeric
Dry Cough	Indicates the severity of dry cough.	Numeric
Snoring	Indicates the presence or severity of snoring.	Numeric
Level	Indicates the lung cancer severity level (Low / Medium / High).	Nominal

Before training the machine learning models, the class distribution of both datasets was analyzed to determine whether class imbalance was present. The lung cancer prediction dataset was examined by comparing the number of positive and negative cases, while the severity prediction dataset was analyzed across the three severity classes (Low, Medium, and High). Understanding the class distribution is important because highly imbalanced datasets may bias the model towards the majority class and affect prediction performance. The observed class distributions of the two datasets are presented in Figures 1 and 2.

B. Data Preprocessing

Data preprocessing is an essential step in developing an accurate and reliable machine learning model. The quality of the input data directly influences the performance of the prediction models. Therefore, the collected datasets are preprocessed before training the machine learning algorithms to ensure consistency and improve prediction accuracy.

Initially, both datasets are examined for missing values and inconsistent records. The datasets are then transformed into a format suitable for machine learning analysis. Numerical attributes, including Age, Blood Pressure, Weight, and Cholesterol, are retained in their original numerical format, while categorical attributes such as Gender, Smoking, Alcohol, Family History, Anemia, Heart Disease, and Diabetes are encoded into numerical values to enable efficient model training. Similarly, in the severity prediction dataset, categorical information is represented using encoded

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 3, July 2026 Impact Factor: 8.2

numerical values according to the predefined dataset labels. These preprocessing steps ensure that all input features can be processed by the implemented machine learning algorithms.

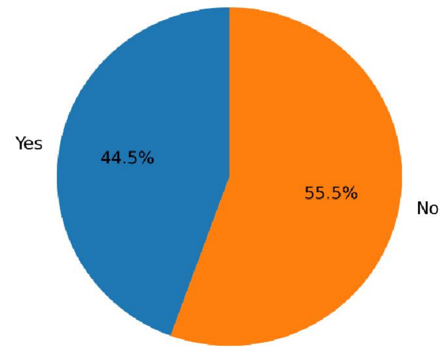


Fig. 1. Class distribution in the Survey Lung Cancer dataset

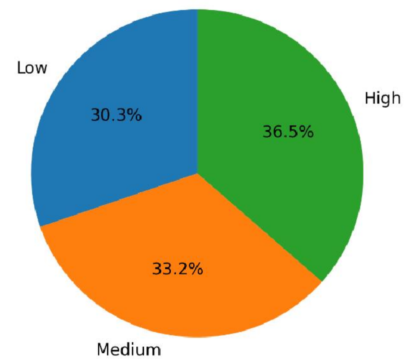


Fig. 2. Class distribution in the Cancer Patient Dataset

For the lung cancer prediction dataset, the target attribute Result is encoded into binary class labels representing the presence or absence of lung cancer. Likewise, in the severity prediction dataset, the target variable is encoded into three severity classes representing Low, Medium, and High severity levels. These encoded target values enable the supervised machine learning models to perform both binary classification for lung cancer prediction and multiclass classification for severity prediction.

After preprocessing, the datasets are verified to ensure that all feature values are properly formatted and ready for model training. The processed datasets are then utilized for training and evaluating the K-Nearest Neighbors (KNN) and Gaussian Naïve Bayes (GNB) algorithms, which are subsequently integrated into the proposed web-based application for lung cancer and severity prediction.

Data Splitting

In the field of machine learning, it is customary to divide the dataset into separate sets before implementing classification models. In this study, an 80-20 splitting method is employed for the two datasets. This approach involved partitioning the data into two sets, with 80% allocated for training the model and the remaining 20% reserved for testing its performance.

C. Classification models

1. K-Nearest Neighbors (KNN): K-Nearest Neighbors (KNN) is a supervised machine learning algorithm widely used for classification tasks. The algorithm classifies a new sample by identifying the K nearest training instances based on a distance metric, commonly Euclidean distance. The class label that appears most frequently among the nearest

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, July 2026
Impact Factor: 8.2

neighbors is assigned to the new sample. Owing to its simplicity and effectiveness, KNN is employed in this study for both lung cancer prediction and severity prediction. The algorithm achieved the highest prediction accuracy among the implemented models and was integrated into the proposed web application.

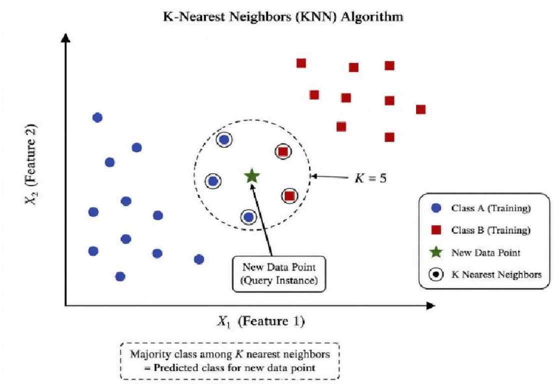


Fig. 3 KNN Illustration

2. Gaussian Naïve Bayes (GNB): Gaussian Naïve Bayes (GNB) is a probabilistic supervised machine learning algorithm based on Bayes' theorem. It assumes that the input features are conditionally independent and follow a Gaussian (normal) distribution. The algorithm calculates the probability of each class based on the feature values and assigns the class with the highest posterior probability. In this study, Gaussian Naïve Bayes is implemented for both lung cancer prediction and severity prediction and is evaluated alongside the KNN model to compare prediction performance.

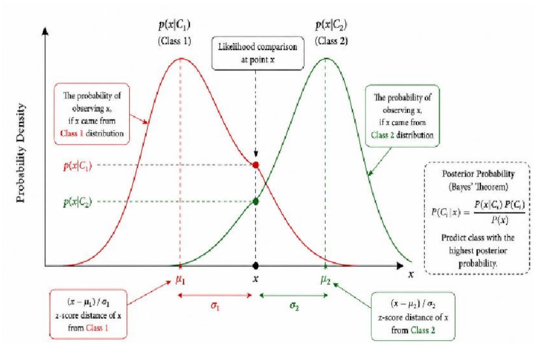
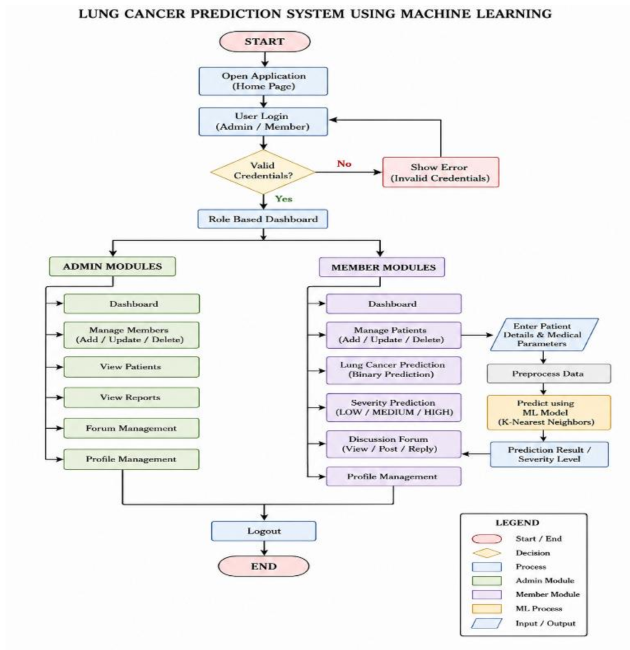


Fig. 4 Gaussian Naïve Bayes Illustration

E. Web Application Development

After training and evaluating the machine learning models, the best-performing model is integrated into a web-based application developed using the Flask framework. The application provides functionalities including user authentication, patient registration, lung cancer prediction, severity prediction, prediction history management, and report generation. Patient information entered through the web interface is processed by the trained machine learning model, and the prediction results are displayed in real time. A MySQL database is used to store patient records, prediction history, and user information, enabling efficient data management within the application.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, July 2026
Impact Factor: 8.2



IV. RESULTS AND DISCUSSION

This section presents the performance evaluation of the proposed lung cancer prediction system using two supervised machine learning algorithms, namely K-Nearest Neighbors (KNN) and Gaussian Naïve Bayes (GNB). The models were trained and evaluated using an 80:20 train-test split on the lung cancer prediction dataset. Their performance was assessed using standard classification metrics, including accuracy, precision, recall, F1-score, confusion matrix, training time, and prediction time. In addition, the KNN classifier was evaluated on the lung cancer severity dataset to classify patients into High, Medium, and Low severity levels.

A. Performance Evaluation of Lung Cancer Prediction Models

The experimental results obtained from the binary lung cancer prediction dataset are summarized in Table III. Both KNN and Gaussian Naïve Bayes demonstrated excellent classification performance. However, the KNN classifier consistently achieved higher evaluation metrics than the Gaussian Naïve Bayes classifier.

TABLE III Performance Comparison of Lung Cancer Prediction Models

Performance Metric	KNN	Gaussian Naïve Bayes
Accuracy (%)	98.43	97.00
Precision (%)	98.46	97.01
Recall (%)	98.43	97.00
F1-Score (%)	98.43	97.00
Training Time (s)	0.0026	0.0025
Prediction Time (s)	0.0215	0.0005
Total Execution Time(s)	0.4375	0.5647

As shown in Table II, the KNN classifier achieved an overall accuracy of 98.43%, outperforming the Gaussian Naïve Bayes classifier, which achieved 97.00% accuracy. Similarly, KNN recorded higher precision, recall, and F1-score values, indicating superior predictive capability. Although the Gaussian Naïve Bayes model required less prediction time due to its probabilistic nature, KNN produced more accurate predictions, making it the better-performing classifier for the proposed system.

ISSN: 2581-9429
Volume 6, Issue 3, July 2026
Impact Factor: 8.2

The confusion matrix of the KNN classifier is presented in Fig. II. The model correctly classified 388 negative and 301 positive cases while misclassifying only 11 samples out of 700 testing instances.

B. Discussion

The experimental evaluation demonstrates that the proposed machine learning-based lung cancer prediction system provides highly accurate classification results. Among the evaluated models, the K-Nearest Neighbors (KNN) classifier consistently outperformed the Gaussian Naïve Bayes (GNB) classifier in terms of accuracy, precision, recall, and F1-score for binary lung cancer prediction. Although Gaussian Naïve Bayes required slightly less prediction time, the higher predictive performance of KNN makes it more suitable for the proposed application.

For lung cancer severity prediction, the KNN classifier achieved perfect classification on the testing dataset, correctly identifying all samples without any misclassification. This result indicates that the selected features in the severity dataset provide strong discriminatory information for classifying patients into High, Medium, and Low severity categories within the scope of this study.

Overall, the experimental results demonstrate that the proposed system effectively predicts both lung cancer occurrence and severity. The trained KNN model was therefore integrated into the Flask-based web application to provide real-time prediction, severity assessment, and report generation, demonstrating the practical applicability of the proposed system in supporting clinical decision-making.

V. CONCLUSION

This study presented a machine learning-based web application for predicting lung cancer and its severity using clinical data. The proposed system employed K-Nearest Neighbors (KNN) and Gaussian Naïve Bayes (GNB) for lung cancer prediction and KNN for severity prediction. Experimental results showed that the KNN model achieved the highest prediction accuracy of 98.43%, while the severity prediction model achieved 100% accuracy on the testing dataset. The developed Flask-based application provides an efficient and user-friendly platform for real-time prediction and severity assessment. Future work may involve evaluating the system on larger and more diverse datasets and exploring additional machine learning techniques to further improve its performance.

REFERENCES

[1] J. S. Muhammad, "Optimizing Lung Cancer Prediction through Feature Selection-Based Machine Learning Models," 2024.

[2] "Prediction of Lung Cancer Using Machine Learning Techniques and Their Comparative Analysis," Proc. IEEE Int. Conf. on Reliability, Infocom Technologies and Optimization (ICRITO), 2022.

[3] "Early Detection of Lung Cancer Using Machine Learning Technique," Proc. IEEE Int. Conf. on Computational Intelligence and Computing Applications (ICCICA), 2023.

[4] "Early Detection of Lung Cancer Using Computer Aided Tomography Images," Proc. IEEE Int. Conf. on Computing, Communication and Networking Technologies (ICCCNT), 2021.

[5] "Detection of Lung Cancer in CT Images Using Image Processing," Proc. IEEE Int. Conf. on Computer Communication and Informatics (ICCCI), 2019.

[6] "Lung Cancer Detection Based on CT-Scan Images with GLCM and Support Vector Machine," IEEE, 2020.

[7] "Leveraging DenseNet and Genetic Algorithms for Lung Cancer Severity Classification," Proc. IEEE Int. Conf., 2023.

[8] "Real-Time Lung Cancer Detection Using GAN-Enhanced LDCT Imaging," Proc. IEEE Int. Conf., 2024.

[9] "Predicting Lung Disease Severity via Image-Based Air Quality Index Analysis," Proc. IEEE Int. Conf., 2024.

[10] "Predicting Lung Cancer Disease Using Optimized Weighting-Based Enhanced Neural Network Classification," Proc. IEEE Int. Conf., 2024.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, July 2026
Impact Factor: 8.2

[11] Survey Lung Cancer Dataset, Kaggle. Available: https://www.kaggle.com/datasets/ankushpanday1/lung-cancer-prediction
[12] Cancer Patients and Air Pollution Dataset, Kaggle. Available: https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer
[13] Python Software Foundation, Python Documentation. Available: https://docs.python.org/3/
[14] Pallets Project, Flask Documentation. Available: https://flask.palletsprojects.com/
[15] Oracle Corporation, MySQL Documentation. Available: https://dev.mysql.com/doc/