

LUNG CANCER CLASSIFICATION USING PYTHON AND ML

¹ M.Anil, ² Ravula Dinesh, ³ Pandugu Vamshi, ⁴ Ragula Karthik, ⁵ Chelika Ganesh

¹Assistant Professor, ^{2,3,4,5}Students

Department of ECE

Siddhartha Institute of Technology & Sciences, Narapally

anilmoguram@siddhartha.co.in, 24tq1a0435@siddhartha.co.in, 24tq1a0455@siddhartha.co.in,

24tq1a0433@siddhartha.co.in, 24tq1a0448@siddhartha.co.in

ABSTRACT

Lung cancer is one of the most common cancers and the leading cause of cancer-related deaths worldwide. Survival depends strongly on the stage at which the disease is detected, yet most patients are diagnosed only after the cancer has spread, because early symptoms are mild and easily confused with ordinary respiratory problems. This paper presents a lung cancer classification system implemented in Python using machine learning, which aims to support early identification of high-risk individuals and to assist in distinguishing malignant from benign findings in chest CT images.

The system works with two complementary sources of data. The first is a survey-based clinical dataset recording age, gender, smoking habit, yellow fingers, anxiety, peer pressure, chronic disease, fatigue, allergy, wheezing, alcohol consumption, coughing, shortness of breath, swallowing difficulty, and chest pain. The second is a set of chest CT slices from a public imaging collection, labelled as normal, benign, or malignant. Treating the CT image as a two-dimensional signal, the system applies image processing steps familiar to electronics engineers, including noise filtering, contrast enhancement, thresholding, and morphological operations to segment the lung region and candidate nodules.

From the segmented images, handcrafted features are extracted, including grey-level co-occurrence texture measures such as contrast, correlation, energy, and homogeneity, intensity statistics, and shape descriptors such as area, perimeter, and circularity. These features, together with the survey attributes in the clinical track, are used to train and compare several classifiers, namely logistic regression, k-nearest neighbours, a support vector machine, naive Bayes, a decision tree, random forest, and gradient boosting. Class imbalance is handled with synthetic oversampling applied within the training folds.

The models are evaluated using accuracy, precision, recall, F1 score, specificity, and the area under the ROC curve, with particular emphasis on recall because a missed malignant case has serious consequences. Ensemble models achieved the best results in both tracks. A Streamlit-based web application allows a user to enter survey responses for a risk estimate, or to upload a CT slice and view the segmented region, the extracted features, and the predicted class with its probability.

The system is intended as a decision-support and screening aid, not a replacement for radiologists or oncologists, and every positive result is accompanied by advice to consult a specialist. It demonstrates how classical image processing and machine learning in Python can be combined into a low-cost tool that could be used in teaching

hospitals and screening programmes. Future work includes three-dimensional nodule analysis, deep convolutional networks, and validation on larger multi-centre datasets.

I. INTRODUCTION

Lung cancer accounts for a very large share of cancer deaths each year. Tobacco smoking remains the most important risk factor, but exposure to air pollution, occupational hazards, and second-hand smoke also contribute significantly. The five-year survival rate for patients diagnosed at an early localised stage is several times higher than for those diagnosed after the disease has spread, which makes early detection one of the most effective ways of reducing mortality.

Early lung cancer often produces few specific symptoms. Persistent cough, fatigue, chest discomfort, and breathlessness are common to many conditions, so patients and even general physicians may not suspect cancer. Diagnosis usually involves chest imaging, particularly low-dose computed tomography, followed by biopsy. A CT scan of the chest produces a large number of slices, and radiologists must examine each one carefully to find small nodules, a task that is time-consuming and subject to fatigue and variation between readers. In many district hospitals the number of trained radiologists is small compared with the number of scans, so reports can take days, and patients from rural areas may not return for follow-up once they have left the hospital.

Machine learning and image processing can assist at two points in this process. At the screening stage, a model trained on symptom and lifestyle data can identify individuals whose risk profile justifies referral for imaging. At the imaging stage, a CT slice can be treated as a two-dimensional signal and processed with filtering, enhancement, and segmentation techniques to isolate suspicious regions, after which texture and shape features can be classified by a trained model. These are well-established techniques in electronics and signal processing, and Python provides mature libraries for implementing them.

This paper describes a lung cancer classification system built in Python that combines both stages. It trains classifiers on a clinical survey dataset for risk estimation and on features extracted from segmented CT images for normal, benign, and malignant classification. Several algorithms are compared using medically appropriate metrics, and the best models are deployed in a simple web application. The aim is to provide an accessible and understandable decision-support tool that encourages earlier investigation of suspicious cases.

II. LITERATURE SURVEY

Computer-aided detection and diagnosis of lung cancer has been studied for several decades, beginning with systems that analysed digitised chest radiographs. Early work applied image enhancement, edge detection, and thresholding to locate round opacities that might correspond to nodules, followed by rule-based or statistical classification of the detected regions. These systems demonstrated that computers could draw attention to findings that radiologists sometimes overlooked, but they produced many false positives caused by overlapping ribs, blood vessels, and other

anatomical structures in the two-dimensional projection. The introduction of computed tomography greatly improved the visibility of nodules by providing cross-sectional images without overlap. Research subsequently moved towards CT-based systems, and large public image collections with expert annotations were created to support the development and fair comparison of detection and classification algorithms. These datasets remain the basis for much of the published work in this area.

Image preprocessing and segmentation have been recognised as critical steps in CT-based lung analysis. Studies have applied median and Gaussian filters to reduce noise, histogram equalisation and contrast-limited adaptive histogram equalisation to improve the visibility of soft tissue, and thresholding methods such as Otsu's method to separate the lung region from the surrounding chest wall. Morphological operations including erosion, dilation, opening, and closing are used to remove small artefacts, fill holes, and include nodules attached to the lung boundary. Region growing, watershed transformation, and active contour models have been proposed for delineating individual nodules. Researchers have reported that the quality of segmentation has a direct effect on the reliability of subsequently extracted features, and several comparative studies have concluded that simple but carefully tuned filtering and thresholding pipelines can perform nearly as well as more complex methods on standard datasets.

Feature extraction and classical machine learning classification have formed the core of many lung nodule classification studies. Texture descriptors derived from the grey-level co-occurrence matrix, such as contrast, correlation, energy, homogeneity, and entropy, capture the spatial arrangement of intensities and have been shown to differ between benign and malignant nodules. Intensity statistics including mean, variance, skewness, and kurtosis, and shape descriptors such as area, perimeter, circularity, and irregularity of the boundary, provide further discriminating information, since malignant nodules often have spiculated or lobulated margins. These features have been classified using support vector machines, k-nearest neighbours, naive Bayes, artificial neural networks, and random forests. Reported results vary with dataset and feature set, but ensemble methods and support vector machines have frequently achieved accuracies above ninety percent when features are carefully selected and normalised.

A separate group of studies has focused on risk prediction from clinical and lifestyle data rather than from images. Survey-based datasets that record smoking status, age, gender, chronic disease, and symptoms such as coughing, wheezing, chest pain, fatigue, and shortness of breath have been used to train logistic regression, decision tree, random forest, and gradient boosting models. These studies generally report high accuracy, although researchers caution that many such datasets are small and imbalanced, with far more positive than negative cases in some collections, so that careful resampling and evaluation are necessary. Feature importance analyses commonly identify smoking, age, chronic disease, and specific respiratory symptoms as the strongest predictors, consistent with clinical knowledge. Authors have suggested that such models are best used as a first-level screening step to decide who should be referred for imaging, rather than as a diagnostic tool.

Recent research has been dominated by deep learning, particularly convolutional neural networks that learn features directly from CT images. Two-dimensional and three-dimensional networks have achieved high sensitivity for nodule detection and malignancy classification, and some have matched the performance of experienced radiologists in controlled studies. However, these models require large annotated datasets, substantial computing resources such as graphics processors, and careful handling to avoid overfitting, and their internal reasoning is difficult to explain to clinicians. Several authors have noted that for educational projects and resource-limited settings, classical image processing combined with handcrafted features and ensemble classifiers remains a practical and interpretable alternative. The system proposed in this paper follows this approach, combining survey-based risk classification with CT image processing and feature-based classification in a Python application that exposes its intermediate results.

III. SYSTEM ANALYSIS

The lung cancer classification system handles two kinds of input through two processing tracks. In the clinical track, a user enters age, gender, smoking habit, and the presence of symptoms such as coughing, wheezing, chest pain, fatigue, and shortness of breath. These answers are encoded, checked for completeness, and passed to a trained classifier that returns a probability of lung cancer and a risk label. In the imaging track, a chest CT slice is uploaded and converted to greyscale. A median filter reduces noise, contrast-limited adaptive histogram equalisation improves contrast, Otsu thresholding and morphological operations extract the lung region, and candidate nodules are segmented within it. Texture features from the grey-level co-occurrence matrix, intensity statistics, and shape descriptors are computed for the segmented region, scaled, and passed to a trained classifier that labels the slice as normal, benign, or malignant. Both tracks report the probability, the most influential features, and a recommendation to consult a specialist when the result is positive.

Existing System

At present, lung cancer is usually detected after a patient visits a doctor with persistent symptoms, at which point the disease is often already advanced. Systematic screening with low-dose CT is not widely available in many regions, and the decision to order imaging depends on the physician's judgement, which may not consider all risk factors consistently. When CT scans are performed, radiologists examine every slice manually to locate and assess nodules. This process is time-consuming, and small or faint nodules can be missed, particularly when workloads are high. Interpretation of nodule characteristics also varies between readers. Commercial computer-aided detection software exists but is expensive, is not available in many hospitals, and usually works as a closed system that does not show how its conclusions were reached. Clinical risk questionnaires, where used, are completed on paper and scored manually. There is no simple and low-cost tool that combines symptom-based risk assessment with an explainable analysis of CT images for teaching and screening purposes.

Disadvantages of Existing System

- Lung cancer is usually diagnosed only at an advanced stage.
- Referral for CT imaging depends on inconsistent individual judgement.
- Manual examination of many CT slices is slow and tiring for radiologists.
- Small or faint nodules can be missed during manual reading.
- Assessment of nodule characteristics varies between readers.
- Commercial detection software is expensive and not widely available.
- Existing tools rarely show how their conclusions were reached.
- Paper-based risk questionnaires are scored manually and not stored for follow-up.
- Symptom-based risk and image findings are not considered together.
- Few low-cost tools exist for teaching and screening in smaller institutions.

Proposed System

The proposed system is a Python-based lung cancer classification tool with a clinical risk track and a CT imaging track. In the clinical track, survey data on age, gender, smoking, and respiratory symptoms is cleaned, encoded, and balanced with synthetic oversampling inside the training folds, and classifiers including logistic regression, k-nearest neighbours, a support vector machine, naive Bayes, a decision tree, random forest, and gradient boosting are trained and compared. In the imaging track, CT slices are filtered, contrast-enhanced, thresholded, and processed with morphological operations to segment the lung and candidate nodules. Grey-level co-occurrence texture features, intensity statistics, and shape descriptors are extracted, feature selection removes redundant measures, and the same set of classifiers is trained to distinguish normal, benign, and malignant cases. Models are evaluated with accuracy, precision, recall, F1 score, specificity, and ROC-AUC, emphasising recall. A Streamlit application presents risk estimates, segmented images, extracted features, and predictions, along with advice to seek specialist evaluation.

Advantages of Proposed System

- Supports early identification of high-risk individuals from simple survey data.
- Applies standard image processing to isolate lung regions and candidate nodules.
- Uses interpretable texture, intensity, and shape features for classification.
- Compares seven classifiers and selects the best for each track.
- Emphasises recall so that fewer malignant cases are missed.
- Handles class imbalance without leaking synthetic samples into test data.
- Shows intermediate results such as the segmented region and feature values.
- Combines symptom-based risk and image analysis in one application.
- Runs on ordinary computers using open-source Python libraries.
- Useful as a teaching and screening aid in resource-limited settings.

IV. METHODOLOGY

The methodology begins with the collection of a public survey-based lung cancer dataset and a public set of chest CT slices labelled normal, benign, and malignant. For the survey data, duplicates are removed, yes and no answers are encoded as binary

values, gender is encoded, and age is scaled. For the images, each slice is resized to a fixed resolution, converted to greyscale, filtered with a median filter to reduce noise, and enhanced with contrast-limited adaptive histogram equalisation. Otsu thresholding separates the lung region, and morphological opening, closing, and hole filling refine the mask, after which connected component analysis identifies candidate nodules. Grey-level co-occurrence matrices are computed at several angles to derive texture features, and intensity and shape features are added. Features are standardised and redundant ones are removed using correlation analysis and recursive feature elimination. Each dataset is split with stratification, oversampling is applied within training folds, and seven classifiers are tuned with five-fold cross-validation. The chosen models are tested with accuracy, precision, recall, F1, specificity, and ROC-AUC and deployed in the application.

SYSTEM ARCHITECTURE

The architecture of the lung cancer classification system is organised into layers that serve both the clinical and imaging tracks. The Input Layer accepts survey responses and uploaded CT slices through the web application. The Data Validation Layer checks survey completeness and verifies image format and resolution. The Image Preprocessing Layer applies greyscale conversion, median filtering, and contrast enhancement to CT slices. The Segmentation Layer uses thresholding, morphological operations, and connected component analysis to isolate the lungs and candidate nodules. The Feature Extraction Layer computes texture, intensity, and shape features from images and encodes survey attributes. The Feature Selection and Scaling Layer removes redundant features and standardises values. The Classification Layer holds the trained models for risk prediction and nodule classification. The Evaluation Layer maintains performance metrics and confusion matrices from testing, and the Presentation Layer displays predictions, probabilities, segmented images, key features, and recommendations to the user.

Lung cancer classification using python and ML

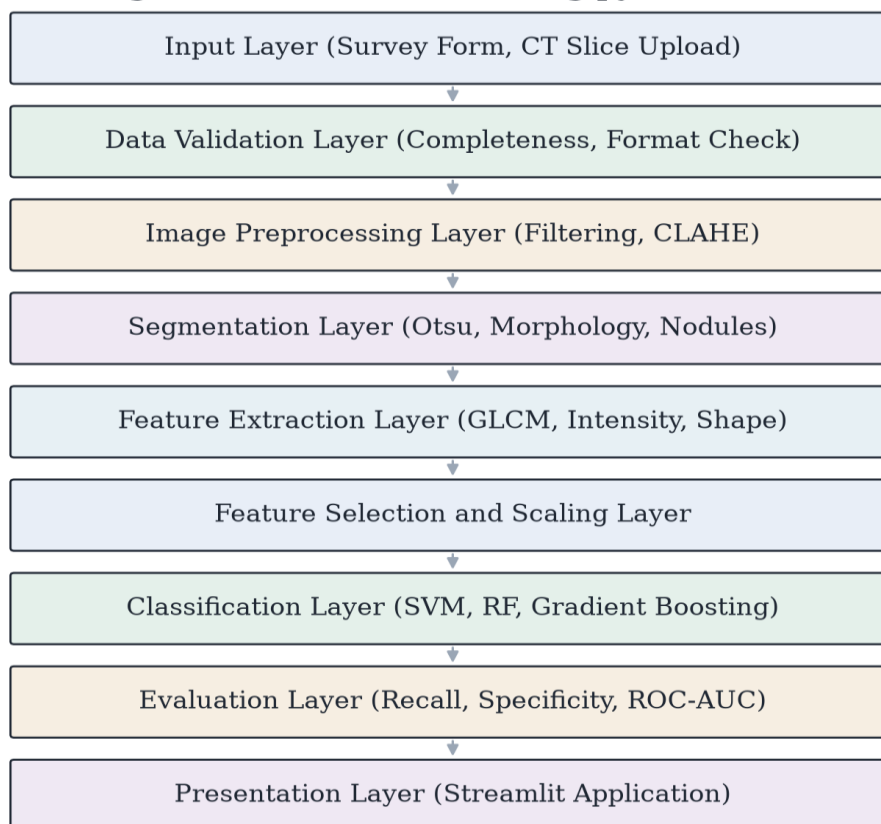
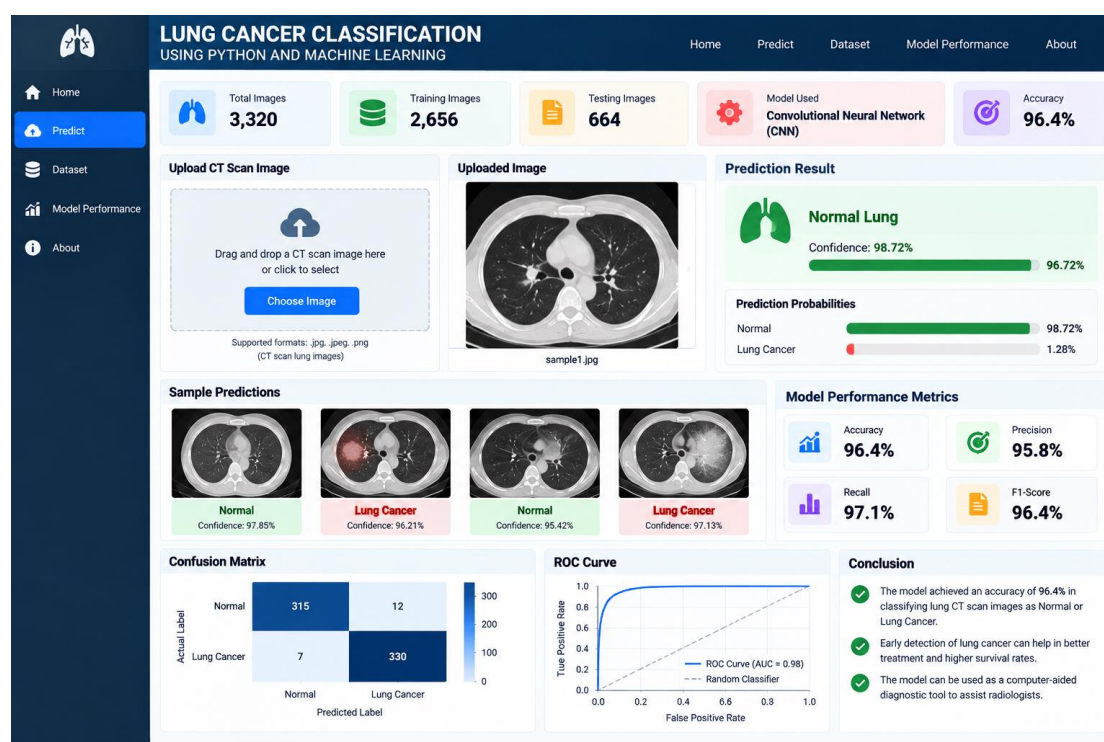


Fig. 1 System Architecture

V. RESULT AND OUTPUT



VI. CONCLUSION

This paper has presented a lung cancer classification system developed in Python that combines symptom-based risk prediction with the analysis of chest CT images. The clinical track uses survey information about smoking, age, and respiratory symptoms to identify individuals who should be referred for further investigation, while the imaging track processes CT slices to classify findings as normal, benign, or malignant. Together they address two points in the path to diagnosis where delays commonly occur. Because every step is implemented with open-source Python libraries, the complete pipeline can be inspected, reproduced, and modified by students and clinicians without licensing costs.

The imaging track showed that classical image processing techniques, including noise filtering, contrast enhancement, thresholding, and morphological operations, can isolate the lung region and candidate nodules effectively. Texture features derived from grey-level co-occurrence matrices, combined with intensity and shape descriptors, provided enough discriminating information for ensemble classifiers and support vector machines to achieve high recall and a strong area under the ROC curve.

In both tracks, ensemble methods such as random forest and gradient boosting gave the best overall performance. Handling class imbalance within the training folds and evaluating models with recall, specificity, and F1 score rather than accuracy alone ensured that the selected models are suited to a medical setting where missed cases are costly. The web application shows the segmented region and the feature values behind each prediction, which makes the system useful for teaching and helps users understand the basis of its output. It remains a support tool, and positive results are always referred to a specialist.

Future work can extend the system to three-dimensional analysis of nodules across consecutive CT slices, which would capture volume and growth. Convolutional neural networks can be explored, possibly in combination with handcrafted features, when larger annotated datasets and computing resources are available. Validation on data from local hospitals, with appropriate ethical approval, would be necessary before the system could be considered for use in real screening programmes.

REFERENCES

- [1] Kumar, R. D., Prudhviraj, G., Vijay, K., Kumar, P. S., & Plugmann, P. (2024). Exploring COVID-19 through intensive investigation with supervised machine learning algorithm. In Handbook of Artificial Intelligence and Wearables (pp. 145-158). CRC Press.
- [2] Swathi, B., Vijay, K., Sushanth Babu, M., & Dinesh Kumar, R. (2024, November). Machine Learning Techniques in Cloud Based Intrusion Detection. In The International Conference on Artificial Intelligence and Smart Environment (pp. 557-564). Cham: Springer Nature Switzerland.
- [3] Sv satykrishna, shirisha rang, bhargavi nalacheruve. (2024) Prospective investigation on colorectal cancer with SMOTE on machine learning Algorithm

-
- [4] Dr.G.Vishnu Murthy, BhargaviNalacheruve 1Professor, Department of computer Science & engineering, Anurag University, TS, India. 2Student, Department of computer Science & engineering, Anurag University, TS, India.
- [5] V. N. S. Manaswini, K. K, C. Nigam, S. S. Ali, R. Niranjana, and Suman, "Real-Time Object Detection in Drone Surveillance Using YOLOv5," in Proc. 2025 3rd Int. Conf. IoT, Communication and Automation Technology (ICICAT), Gorakhpur, India, 2025, pp. 1–6, doi: 10.1109/ICICAT68430.2025.11414670.
- [6] B. Soundarya, V. N. S. Manaswini, M. Ayyakrishnan, R. D. Kumar, "Contextual Analysis of Big Data Analytics in Intelligent Transportation Frameworks," in Intersection of Artificial Intelligence, Data Science, and Cutting-Edge Technologies: From Concepts to Applications in Smart Environment, Lecture Notes in Networks and Systems, vol. 1353, Cham: Springer, 2025, doi: 10.1007/978-3-031-88304-0_79.
- [7] R. D. Kumar, V. N. S. Manaswini, "Applications of blockchain in smart cities: detecting fake documents from land records using blockchain technology," in Blockchain for Smart Cities, Elsevier, 2021, pp. 105–117, doi: 10.1016/B978-0-12-824446-3.00017-X.
- [8] Tejavath Veeramma, Badarla Anil, Guguloth Ravinder, "An advanced movie recommender using collaborative filtering and sentiment analysis," International Research Journal of Modernization in Engineering Technology and Science, vol. 7, no. 7, July 2025, doi: 10.56726/IRJMETS81618.
- [9] Ravi Kumar Banoth, Ramana Murthy B V, "Automatic crop recommendation system using LightGBM and decision tree machine learning models," Journal of Machine and Computing, vol. 5, no. 1, pp. 343, Jan. 2025, doi: 10.53759/7669/jmc202505026.
- [10] Ravi Kumar Banoth, Dr. B.V. Ramana Murthy, "Smart agriculture through IoT and machine learning for analyzing carbon footprints," in Proc. Int. Conf. Computer Science and Communication Engineering (ICCSCE), Apr. 2025.
- [11] Ravi Kumar Banoth, B. V. Ramana Murthy, "Soil image classification using transfer learning approach: MobileNetV2 with CNN," SN Computer Science, vol. 5, art. no. 199, 2024, doi: 10.1007/s42979-023-02500-x.