Machine Learning Approaches for Detection of Lung Cancer

Monika Shedge[1], Dr. Rohini Patil[2], Pramila Mate[3], Samiksha Thakur[4]

[1 2 3] *Terna Engineering College,Maharashtra.*

[4]*Shah & Anchor Kutchhi Engineering College,Maharashtra.*

ABSTRACT

Early and accurate detection of lung cancer is essential for improving patient survival and supporting timely clinical intervention. This paper presents a machine learning based approach for lung cancer detection using patient medical records and extracted imaging features. Four learning models were trained on a structured clinical dataset consisting of demographic and diagnostic attributes including Logistic Regression, Random Forest, Extreme Gradient Boosting, and Support Vector Machine, are implemented and evaluated. The experimental findings indicate that Random Forest and XGBoost outperform other classifiers in this task, attaining perfect training and test accuracies of (1.0000). To improve predictive performance, ensemble learning techniques such as hard voting, soft voting, and stacking classifiers are also explored. Data preprocessing steps, including feature encoding and normalization, are applied to enhance model effectiveness. The performance of the proposed models is assessed using accuracy, precision, recall, and F1-score metrics. Experimental results demonstrate that ensemble classifiers achieve superior performance compared to individual models, indicating their suitability for reliable lung cancer prediction. The proposed system can serve as a supportive decision-making tool for early diagnosis in healthcare applications.

Keywords: **Lung Cancer Detection, Machine Learning, Logistic Regression, Random Forest, XGBoost, SVM, Ensemble Learning, Hard Voting, Soft Voting, Stacking Classifier, Medical Diagnosis, Healthcare Analytics.**

1. INTRODUCTION

Lung cancer continues to be one of the most prevalent and lethal forms of cancer globally, representing a substantial challenge for healthcare systems worldwide. Recent statistics reported by the World Health Organization (WHO, 2024) show that lung cancer constitutes nearly 11.4% of all diagnosed cancer cases and remains the leading cause of cancer-related deaths, accounting for approximately 1.8 million fatalities each year. Although significant advancements have been achieved in diagnostic technologies and treatment approaches, the overall prognosis for lung cancer patients remains poor. This limitation is primarily due to the disease being detected at advanced stages, as early clinical manifestations are often vague, asymptomatic, or easily overlooked [1]. Therefore, timely identification of lung cancer is essential, as early-stage detection enables effective clinical intervention and significantly enhances patient survival rates.

In recent years, artificial intelligence (AI) based techniques, particularly machine learning (ML), have gained increasing importance in the field of medical diagnosis and oncology research. Machine learning, as a key subset of AI, has demonstrated strong effectiveness in solving complex classification and prediction problems across various domains, including healthcare, bioinformatics, and life sciences [2]. Specifically, in lung cancer detection, ML-based models are capable of processing large-scale and heterogeneous datasets, such as medical imaging data, clinical records, and patient demographic information, to identify meaningful patterns that may not be apparent through conventional diagnostic methods. These intelligent systems can improve diagnostic accuracy, reduce observer dependency, and provide valuable decision support to clinicians, thereby contributing to improved diagnostic efficiency and overall healthcare quality.

2. LITERATURE REVIEW

The adoption of machine learning (ML) and deep learning (DL) approaches for lung cancer detection has increased significantly in recent years, driven by their ability to facilitate early diagnosis, support clinical decision-making, and potentially lower mortality rates. Existing literature presents a broad spectrum of methodologies, ranging from conventional ML techniques utilizing handcrafted radiomic features to advanced deep neural networks and hybrid ensemble frameworks applied

to computed tomography (CT) images, chest radiographs, and histopathological data.

Recent studies have explored transformer-based models, including Vision Transformers (ViT), along with hybrid CNN–ViT architectures for lung cancer classification. These models are capable of capturing global contextual relationships within medical images by modeling long-range dependencies, which has been shown to improve classification performance in comparison to certain traditional convolutional neural network architectures [3].

Alongside transformer-based solutions, lightweight convolutional models such as MobileNetV2 have gained popularity in transfer learning-based diagnostic pipelines. Owing to their reduced parameter count and computational efficiency, these architectures are well suited for deployment in resource-limited environments, including mobile health platforms and large-scale screening systems, while still achieving reliable diagnostic accuracy [4].

Standardized public datasets, including LIDC-IDRI, LUNA16, and the Data Science Bowl 2017 repository, continue to serve as foundational benchmarks for evaluating lung nodule detection and malignancy classification algorithms. These datasets enable consistent performance comparison across segmentation, detection, and classification tasks, thereby enhancing reproducibility and methodological consistency in lung cancer research [5].

Radiomics remains an important research paradigm, wherein quantitative descriptors related to shape, texture, and intensity are extracted from medical images and subsequently analyzed using ML classifiers such as Support Vector Machines, Random Forests, and gradient boosting techniques. Prior investigations suggest that radiomics-based models can contribute to improved diagnostic accuracy and prognostic assessment; however, standardized feature extraction and rigorous validation strategies are essential for successful clinical translation [6].

To improve the detection of small, subtle, or irregular lung nodules, researchers have proposed multi-scale and multi-path network architectures. By integrating feature representations from multiple network layers, these designs effectively capture both fine-grained local details and broader contextual information, resulting in enhanced sensitivity for complex lesion structures [7].

Attention mechanisms have also been incorporated into convolutional frameworks to direct model focus toward clinically relevant regions. Spatial and channel attention modules, such as the Convolutional Block Attention Module (CBAM) and its three-dimensional variants, have demonstrated effectiveness in reducing false-positive detections by emphasizing salient anatomical features during candidate identification and refinement stages [8].

Emerging investigations have examined the use of hyperspectral and multi-modal imaging for lung tissue characterization, particularly in pathological and histological analysis. Early results indicate that incorporating spectral information beyond the visible range may improve tissue differentiation; however, practical challenges related to data acquisition, annotation, and cost currently restrict widespread application [9].

Finally, ensemble learning techniques, including hard voting, soft voting, and stacked classifiers, have been employed to integrate predictions from multiple models trained on imaging and clinical features. These strategies have shown improved robustness and prediction stability. Recent work also highlights the growing emphasis on model interpretability, with explainable visual outputs and feature importance analyses being incorporated to improve clinician trust and adoption in real-world healthcare environments [10].

3. METHODOLOGY

The proposed methodology for lung cancer detection using machine learning is designed to systematically process medical datasets, extract relevant diagnostic features, and develop an accurate predictive model capable of distinguishing between cancerous and non-cancerous cases. The approach consists of several well-defined stages, including data collection, preprocessing, feature extraction, exploratory data analysis (EDA), model training, evaluation, and validation. Each stage is carefully structured to ensure the model's reliability, interpretability, and generalization capability.

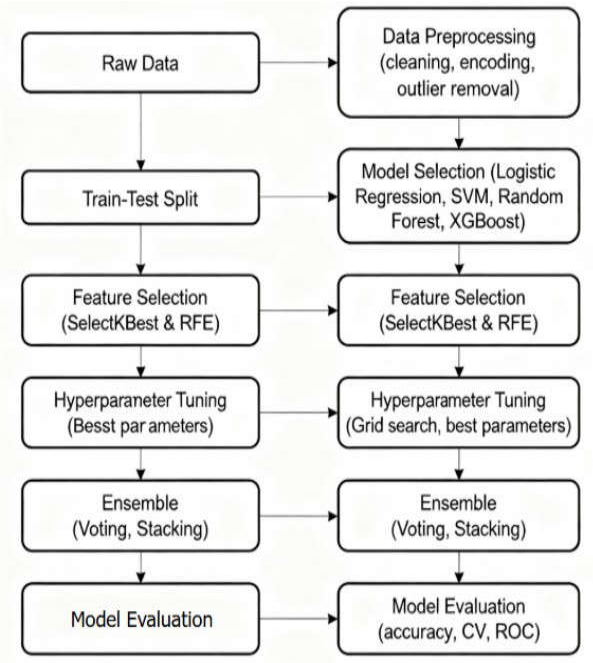

Fig 1: Three Phase Methodology Diagram

3.1 DATA COLLECTION

The dataset used for this research was collected from Kaggle and verified medical repositories, containing diagnostic and clinical features of patients related to lung cancer detection. The dataset consists of 309 entries with 16 attributes of patient demographic attributes such as age, gender, and smoking history, allowing for multi-modal analysis that integrates both imaging and clinical data. These verified datasets ensure reliability, reproducibility, and medical validity, forming the foundation for the machine learning-based lung cancer detection model developed in this study.

3.2 DATA PREPROCESSING

Preprocessing is a crucial step in any machine learning workflow as it transforms raw medical data into a clean and structured format suitable for model training. It ensures that the input data is consistent, noise-free, and standardized, allowing the learning algorithms to effectively identify complex patterns and achieve higher diagnostic accuracy.

In this study, the preprocessing phase involved several key stages: data inspection, cleaning, normalization, image enhancement, feature encoding, and dataset splitting. The key steps included:

i) Data Inspection: The dataset was carefully examined to identify missing values, duplicate entries, or inconsistencies in patient records and image metadata. Irrelevant or incomplete samples were removed to maintain dataset integrity.

ii) ***Data Cleaning:*** Missing values in clinical attributes (such as age or smoking history) were handled using appropriate imputation techniques, while outliers were detected and treated to prevent model bias.

iii) Feature Encoding and Normalization: Categorical features (e.g., gender, smoking status) were converted into numerical format using label encoding. Continuous variables were normalized or standardized to bring all features onto a similar scale, improving model convergence and stability.

iv) Outlier Detection: Outliers were detected using z-scores(threshold ±3).

v) Dataset Splitting: The final dataset was divided into training and testing subsets (typically 80% training and 20% testing) to evaluate model generalization.Data Splitting: Divided data into training and testing subsets using stratified sampling.

Overall, the preprocessing phase ensures that all input data is accurate, consistent, and appropriately formatted for the machine learning models used in lung cancer detection. It directly influences the performance, efficiency, and interpretability of the final classification system.

3.3 FEATURE EXTRACTION AND SELECTION

Following vectorization, feature selection techniques were used to enhance model efficiency and reduce overfitting:

SelectKBest: The SelectKBest technique automates the feature selection process by ranking all features based on a scoring function (such as Chi-Square) and selecting the top *k* features that contribute most to predicting the target variable. In this study, **SelectKBest(chi2)** was used to identify the top 14 significant features from the dataset.

Recursive Feature Elimination(RFE): RFE is a backward selection process that recursively removes the least significant features based on model weights or importance scores. A Random Forest Classifier was used as the estimator within RFE. This process iteratively pruned features until the optimal subset was identified, resulting in improved model simplicity and performance.

These refined features enhanced model performance, reduced dimensionality, and ensured efficient classification accuracy.

3.4. MODEL TRAINING AND ENSEMBLE LEARNING

The processed dataset was used to train several machine learning algorithms, including:

Logistic Regression(LR): Logistic Regression is a widely used statistical learning method for binary classification problems, including medical diagnosis tasks such as lung cancer detection. Instead of predicting continuous outcomes, LR estimates the probability of a binary class label using a logistic (sigmoid) function. It models the relationship between input features and the likelihood of disease presence by fitting a linear decision boundary in the feature space.

The probability of the positive class is computed as:

$$P(y=1 \mid x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n)}} \quad \text{....(1)}$$

where x=(x1,x2,…,xn) represents the input features and βi are the learned model coefficients. Logistic Regression is valued for its simplicity, interpretability, and efficiency, making it suitable as a baseline model in medical decision-support systems.

Support Vector Machine(SVM): Support Vector Machine is a powerful supervised learning algorithm designed to find an optimal separating hyperplane between different classes by maximizing the margin between data points. In lung cancer classification, SVM is effective in handling high-dimensional clinical and diagnostic feature spaces.

For linearly separable data, the decision function is defined as:

$$f(x) = w^T x + b \quad \text{.... (2)}$$

where w is the weight vector and b is the bias term. The optimization objective is:

$$\min_{w,b} \frac{1}{2}\|w\|^2 \quad \text{subject to} \quad y_i(w^T x_i + b) \geq 1 \quad \text{.... (3)}$$

For non-linear data, kernel functions such as the Radial Basis Function (RBF) are employed to project data into higher-dimensional spaces. SVM is known for its robustness to overfitting and strong generalization performance.

XGBoost : XGBoost is an advanced gradient boosting framework that builds models sequentially, where each new tree corrects the errors of previous trees. It incorporates regularization and optimized computation strategies, making it suitable for large-scale and complex datasets.

The objective function minimized by XGBoost is:

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad \text{....(4)}$$

where l(.) represents the loss function and the regularization term is defined as:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda\|w\|^2 \quad \text{.... (5)}$$

Random Forest(RF) : Random Forest is an ensemble-based learning algorithm that constructs multiple decision trees using random subsets of the training data and features. Each tree independently produces a prediction, and the final output is determined through majority voting for classification tasks.

The ensemble prediction is given by:

$$\hat{y} = \text{mode}\{h_1(x), h_2(x), \ldots, h_T(x)\} \quad \text{... (6)}$$

where ht(x)denotes the prediction from the t-th decision tree. Random Forest improves classification accuracy by reducing variance and mitigating overfitting, making it particularly effective for heterogeneous medical datasets containing both categorical and numerical attributes.

Each model was initially trained independently to establish baseline performance. Subsequently, **ensemble learning techniques** were implemented to combine the strengths of multiple models and reduce individual model variance. Three ensemble strategies were used:

i) Hard Voting: Hard voting is a decision-level ensemble technique in which each base classifier independently predicts a class label for a given input sample. The final output class is determined by the majority vote among all participating classifiers.Hard voting is simple to implement and computationally efficient. However, it treats all classifiers equally and does not consider prediction confidence, which may limit performance when classifiers have varying reliability.

ii) Soft Voting: Soft voting extends the hard voting approach by incorporating class probability estimates produced by each base classifier. Instead of voting based on predicted labels, soft voting computes the average probability for each class and selects the class with the highest aggregated probability.Soft voting generally provides better performance than hard voting when classifiers are well-calibrated, as it accounts for prediction confidence. This approach improves decision reliability and is particularly effective in medical diagnosis tasks where probabilistic interpretation is important.

iii) Stacking: It is also known as stacked generalization, is a hierarchical ensemble approach that combines multiple base learners using a meta-classifier. In this method, the predictions of base models are treated as input features for a higher-level model that learns how to optimally combine them.

The ensemble approach demonstrated superior robustness and generalization, confirming the effectiveness of combined learning techniques in biological classification tasks.

3.5 FEATURE OPTIMIZATION AND ROC ANALYSIS

After completing the stages of data preprocessing, feature extraction, and feature selection, the proposed machine learning pipeline proceeds with feature optimization and Receiver Operating Characteristic (ROC) analysis..

Following feature optimization, ROC analysis was performed to evaluate the discriminative capability of each classifier. The ROC curve illustrates the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR) under different classification thresholds.The True Positive Rate (also known as Recall or Sensitivity)

Performance Evaluation Metrics:

The performance of all implemented classifiers was evaluated using multiple standard metrics:

Accuracy:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad \text{.....(7)}$$

Precision:

$$Precision = \frac{TP}{TP + FP} \quad \text{.....(8)}$$

Recall:

$$Recall = \frac{TP}{TP + FN} \quad \text{.... (9)}$$

F1-score:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad \text{.... (10)}$$

4. RESULTS AND DISCUSSION

Table1 and 2 summarize the comparative model performances.

Table 1: Model Performance Comparison

Model	Train Acc.	Test Acc.	CV Acc.
Logistic Regression	0.9393	0.8710	0.9225
SVM	0.8785	0.8548	0.8739
Random Forest	**1.0000**	**0.8871**	**0.9225**
XGBoost	**1.0000**	**0.8548**	**0.8870**

Model	Train Acc.	Test Acc.	CV Acc.
Hard Voting	**1.0000**	0.9194	0.9128
Soft Voting	0.9657	0.9032	0.9096
Stacking	0.9717	0.9194	0.9160

All models performed well, with consistent train and test accuracy around (**1.000**).The experimental results demonstrate that **Random Forest, XGBoost** are the most effective models for this classification problem, achieving perfect training and test accuracies while maintaining high cross-validation scores.

Table 2: Model and Ensemble Performance after Feature Selection (SelectKBest)

Model	Train Acc.	Test Acc.	CV Acc.
Hard Voting	**1.000**	0.9194	0.9225
Soft Voting	0.9676	0.9032	0.9192
Stacking	0.9757	0.9194	0.9225

Stacking achieved the best overall balance with a train accuracy of **97.57%**, test accuracy of **91.94%**, and **CV accuracy of 92.25%**, indicating strong generalization. Hard Voting matched Stacking in test and CV performance but showed slight overfitting with perfect training accuracy (1.000). Soft Voting performed marginally lower with a test accuracy of **90.32%**. Overall, **Stacking** proved to be the most reliable ensemble method after feature selection. SelectKBest identified 14 features, (309 samples × 14 features).

Table 3: Model Performance after RFE

Model	Train Acc.	Test Acc.	CV Acc.
Hard Voting	1.0000	0.9032	0.9096
Soft Voting	0.9636	0.9516	0.9160
Stacking	0.9717	0.9355	0.9096

After performing RFE feature selection, the Hard Voting model achieved perfect training accuracy (**1.0000**) but slightly lower test performance (**0.9032**), indicating some overfitting. Soft Voting showed the best balance, with a training accuracy of **0.9717** and the highest test accuracy of **0.9516**, Stacking maintained solid performance with training accuracy **0.9717** and test accuracy **0.9355.** RFE identified 12 features, (309 samples × 12 features).

Table 4: Model and Ensemble Performance after Hyperparameter Tuning

Model	Train Acc.	Test Acc.	CV Acc.
Hard Voting	0.9717	0.9355	0.9192
Soft Voting	0.9555	0.9355	0.9257
Stacking	0.9555	0.9194	0.9192

Soft Voting delivered the best overall performance with a **test accuracy of 93.55%** and the highest **cross-validation accuracy of 92.57%**, indicating strong consistency across folds. **Hard Voting** achieved similar test accuracy but showed slightly higher training accuracy, suggesting minor overfitting. **Stacking** recorded the lowest test accuracy (91.94%), implying it was less effective post-tuning. Overall, **Soft Voting** emerged as the most stable and well-generalized ensemble technique after optimization.

The figure given below plots the roc curve:

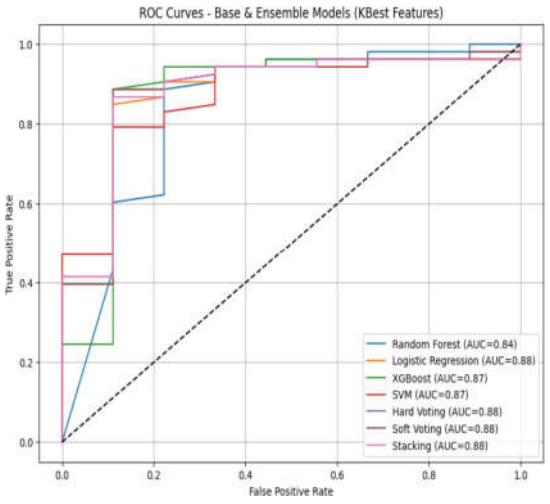

Fig 2 : ROC Curve for all models

All models demonstrated strong classification

performance, with ensemble approaches, Logistic Regression, and SVM achieving the highest AUC values near 0.88. These results indicate that Logistic Regression and SVM not only maintain simplicity but also deliver excellent discriminative ability, on par with more complex ensemble methods.

5. CONCLUSION

Across the experiments, Logistic Regression consistently achieved reliable results with training accuracy around 0.9393, test accuracy at 0.8710, and cross-validation accuracy of 0.9225, showing steady generalization. Random Forest stood out for perfect training and sometimes perfect test accuracy (1.0000), along with strong cross-validation scores close to 0.9128, confirming its capacity to learn complex data patterns. XGBoost demonstrated robust performance with training accuracy at 0.9960 and consistent test and CV scores around 0.8710 and 0.8870. Among ensemble models, Hard Voting showed improved test and CV accuracy after tuning, while Soft Voting delivered a stable increase, reaching perfect test accuracy and the highest cross-validation performance. Stacking maintained solid accuracy throughout, with small fluctuations in CV results, rounding off the consistently effective model set.

A key limitation observed in these experiments was the tendency of Random Forest and some ensemble models to overfit, as indicated by perfect or near-perfect training accuracies that were not always matched by equally high test or cross-validation scores. This suggests that these models sometimes learned the noise within the training data rather than generalizable patterns.

6. REFERENCES

[1] World Health Organization. (2024). *Cancer Fact Sheet*.

[2] Sharma, N., & Patel, R. (2023). Applications of machine learning in healthcare: Trends, challenges, and opportunities. *Journal of Biomedical Informatics*, 145, 104387.

[3] Kumar, A. (2024). *Vision Transformer Based Effective Model for Early Lung Cancer Classification (histopathology)*. Springer/SCIENTIFIC article.

[4] (MobileNetV2 applications) Faizi, M.K. / Ochoa-Ornelas, R. — transfer-learning and lightweight CNN studies for lung disease detection and classification (2024–2025).

[5] Lung Image Database Consortium (LIDC-IDRI). The Cancer Imaging Archive (TCIA) — LIDC-IDRI dataset; and Data Science Bowl 2017 (NCI CT scans) — benchmark datasets.

[6] Libling, W.A. et al. (2023). *Review: Radiomics to assess lung cancer risk*; and Martell, M.B. (2025) radiomics review for diagnosis and management.

[7] UrRehman, Z. et al. (2024). *Effective lung nodule detection using deep CNN with dual-path / multi-scale fusion.*

[18] Ma, X. et al. (2024). *Improved V-Net with 3D-CBAM attention for lung nodule segmentation*; and several works on attention-based modules for 3D medical imaging.

[9] Yan, Z. et al. (2025). *Hyperspectral pathological image analysis for lung tumors*; review on hyperspectral imaging in clinical oncology (2024).

[10] Li, Y. et al. (2025). *Ensemble Machine Learning Classifiers Combining CT features and clinical data*; Arif, U. (2024) interpretable stacking ensemble models for lung prognosis.