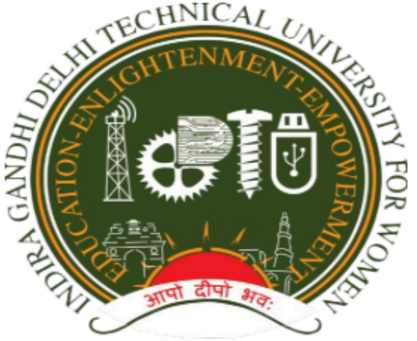


14th International Conference on

Computing, Communication and Sensor Network.CCSN2025

A Machine Learning Approach to Early Lung Cancer Diagnosis

Presented by

Khushi Dutta

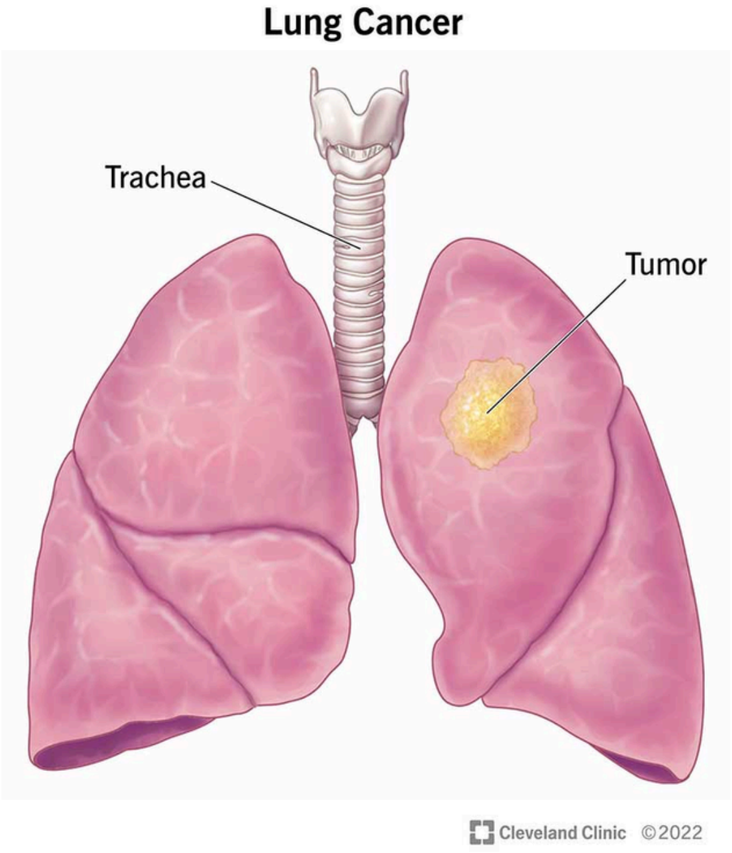
Manya Singh

Abstract

- Lung cancer is the leading cause of cancer-related death worldwide, largely due to late-stage diagnosis and rapid disease progression.
- Early detection is challenging as symptoms often mimic other respiratory illnesses or remain unnoticed until advanced stages.
- The research applies supervised machine learning models such as K-Nearest Neighbors, Naive Bayes, Random Forest, Decision Tree, Support Vector Machine, and Logistic Regression for early and accurate diagnosis using patient clinical and demographic data.
- Results show high accuracy and reliability for AI-driven screening, supporting improved medical decision-making and earlier intervention.
- Persistent challenges include data imbalance and interpretability, though refined data preprocessing and model tuning improve outcomes.

Introduction to Lung Cancer

- Lung cancer is a disease characterized by uncontrolled growth of abnormal cells in the lungs, which can interfere with normal lung function and spread (metastasize) to other parts of the body.
- It is one of the most common and deadly cancers worldwide, primarily caused by factors like tobacco smoking, exposure to air pollution, genetic predispositions, and occupational hazards.
- The two main types are:
- Non-Small Cell Lung Cancer (NSCLC): Accounts for about 85% of cases and grows/spreads more slowly.
- Small Cell Lung Cancer (SCLC): More aggressive and spreads rapidly.
- Common symptoms include persistent cough, chest pain, breathlessness, fatigue, and unexplained weight loss.



Lung Cancer: Global Challenge & Need for Early Detection

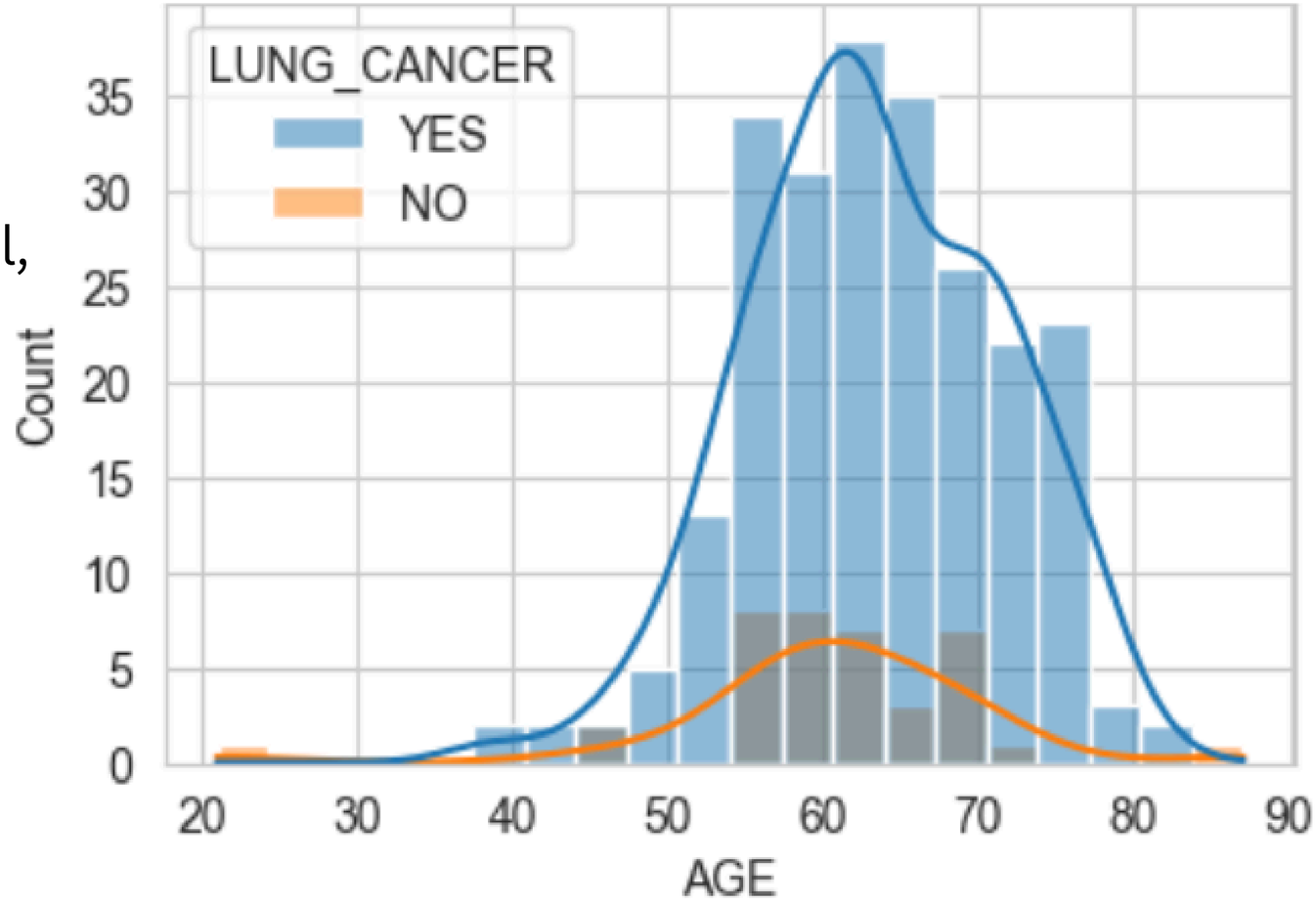
- Lung cancer remains one of the most significant global health challenges, causing approximately 10 million cancer deaths in 2020 alone, with nearly 70% occurring in low- and middle-income countries due to disparities in healthcare access.
- It is the most lethal cancer type, responsible for around 1.8 million deaths, surpassing colorectal, liver, and stomach cancers.
- Early detection of lung cancer is difficult since symptoms can resemble other respiratory conditions or be absent initially, hindering timely diagnosis.
- Effective treatments and better outcomes are possible only if lung cancer is diagnosed at early stages, highlighting the urgency for improved diagnostic tools.
- Tobacco use is the primary risk factor, but environmental and genetic factors (air pollution, occupational hazards, secondhand smoke) also contribute to lung cancer incidence.
- Increasing diagnostic challenges due to growing patient data complexity necessitate automated tools leveraging AI and machine learning for early and accurate detection

Research Gaps

- Traditional lung cancer diagnostic methods often require extensive resources and rely heavily on expert interpretation, which may not always be accessible.
- There is a lack of direct comparison between deep learning models (like CNNs) and traditional machine learning models, despite deep learning's potential.
- Research on hybrid models incorporating GANs, federated learning, transfer learning, and data augmentation in lung cancer detection remains sparse.
- Real-world deployment of AI systems faces challenges such as class imbalance, overfitting, and patient data privacy concerns.

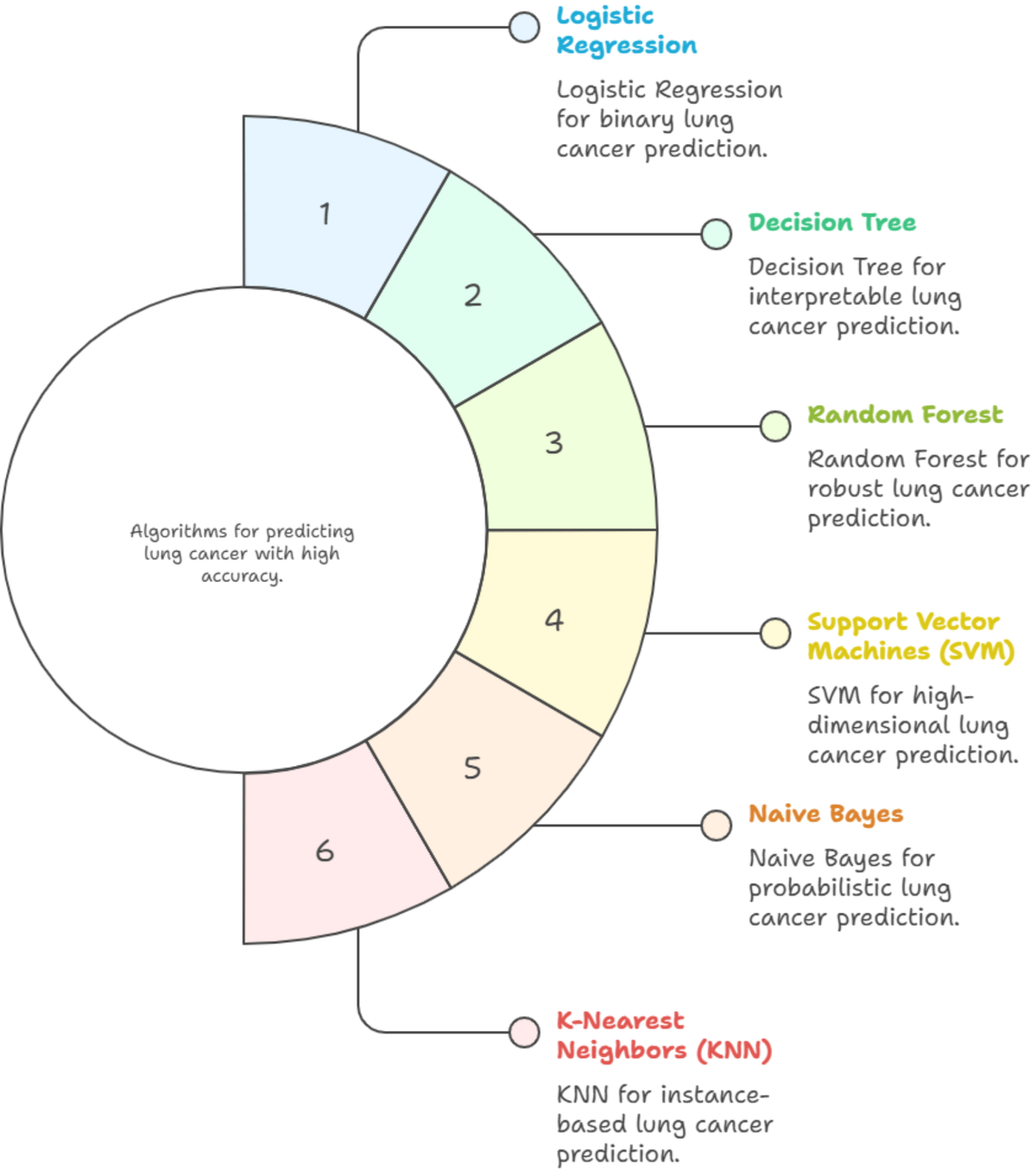
Dataset

- Total Records: 284 unique patient entries after removing 33 duplicates from the original 309 records.
- Attributes: 16 features covering clinical, demographic, and symptom variables:
- Gender (M/F; one-hot encoded)
- Age (numeric)
- Smoking, Anxiety, Yellow Fingers, Chronic Disease, and other symptom/lifestyle factors (binary: 1 = NO, 2 = YES).
- Target Variable: Lung cancer diagnosis (one-hot encoded).

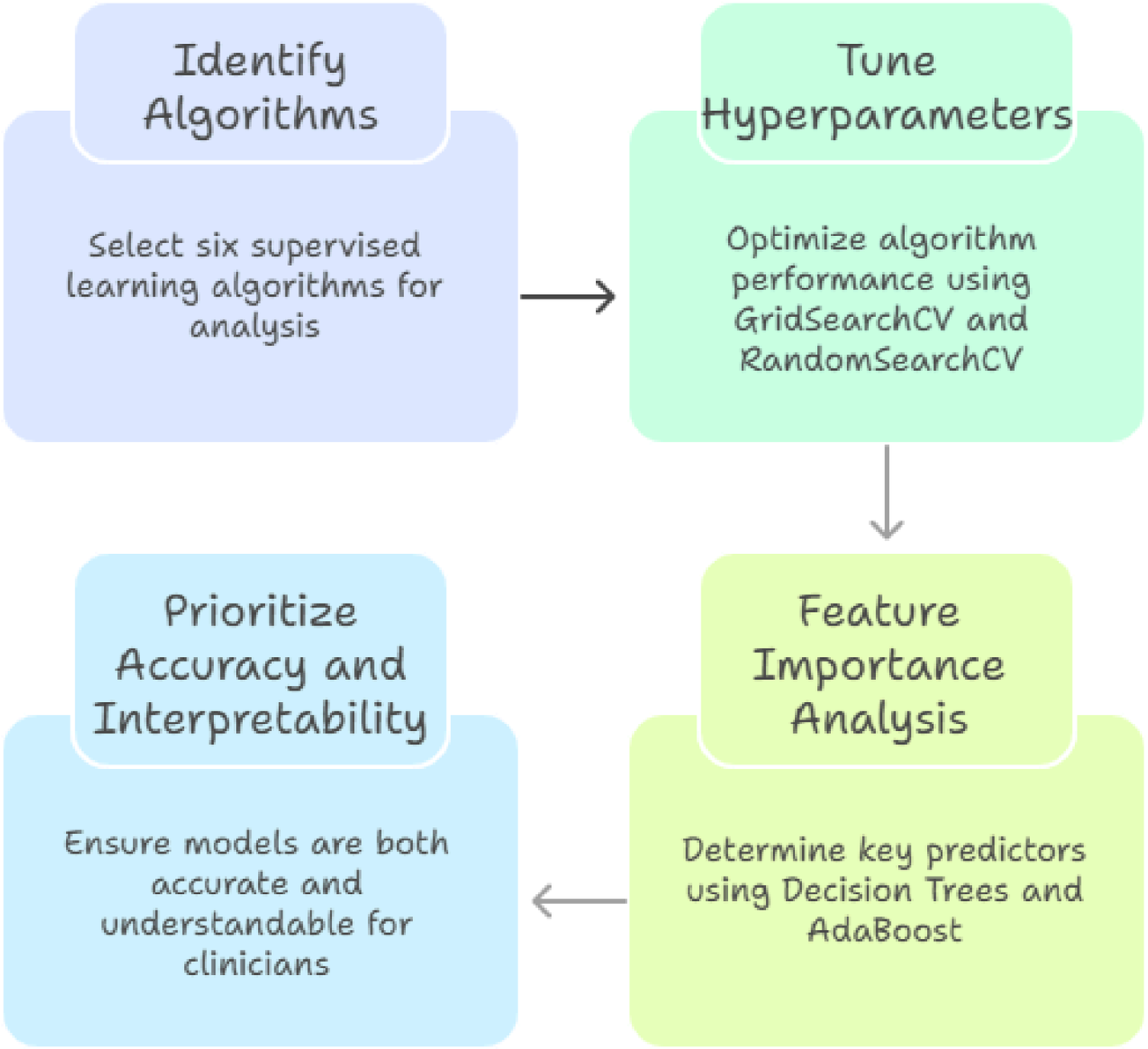


Dataset Prepration & EDA

- Cleaned the dataset by removing duplicate patient records, ensuring every entry counted toward meaningful analysis.
- Transformed categorical details, like gender and diagnosis, using one-hot encoding for compatibility with machine learning models.
- Leveraged Seaborn and Matplotlib to bring data patterns to light, using histograms, count plots, and pair plots to visualize distributions and relationships.
- Data visualization made outliers and trends easy to spot, helping reveal which symptoms and risk factors deserve special attention for early diagnosis.

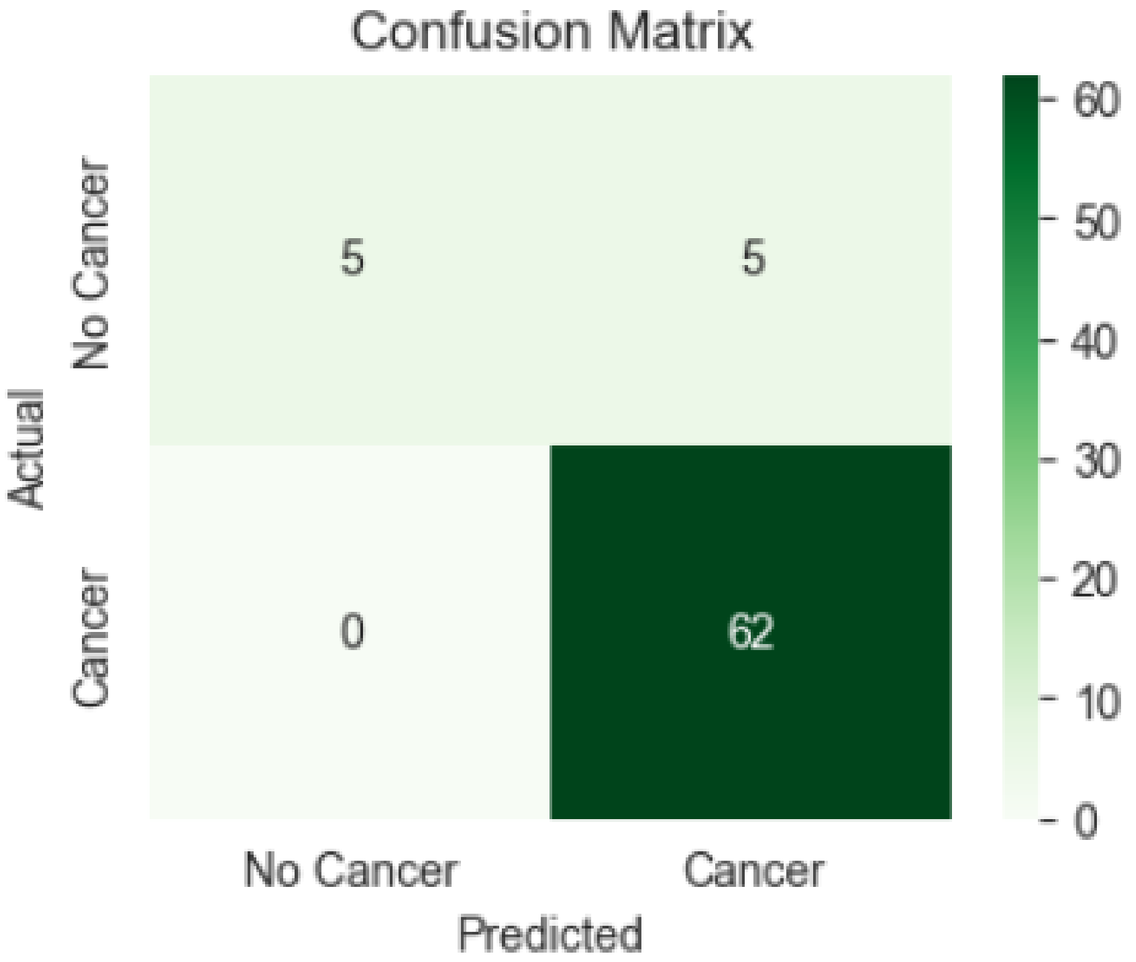


- Evaluated six popular supervised algorithms for lung cancer prediction: Logistic Regression, Decision Tree, Random Forest, SVM, Naive Bayes, and K-Nearest Neighbors (KNN) known for their diverse strengths in handling clinical data.
- Each algorithm brings a unique approach from Logistic Regression's simplicity and interpretability to Random Forest's ensemble robustness and SVM's margin-based classification.
- Comparing these models helps identify which best captures complex patterns in patient data to accurately predict lung cancer risk.


Identify Algorithms
Select six supervised learning algorithms for analysis
Tune Hyperparameters
Optimize algorithm performance using GridSearchCV and RandomSearchCV
Feature Importance Analysis
Determine key predictors using Decision Trees and AdaBoost
Prioritize Accuracy and Interpretability
Ensure models are both accurate and understandable for clinicians

Performance Metrics & Results

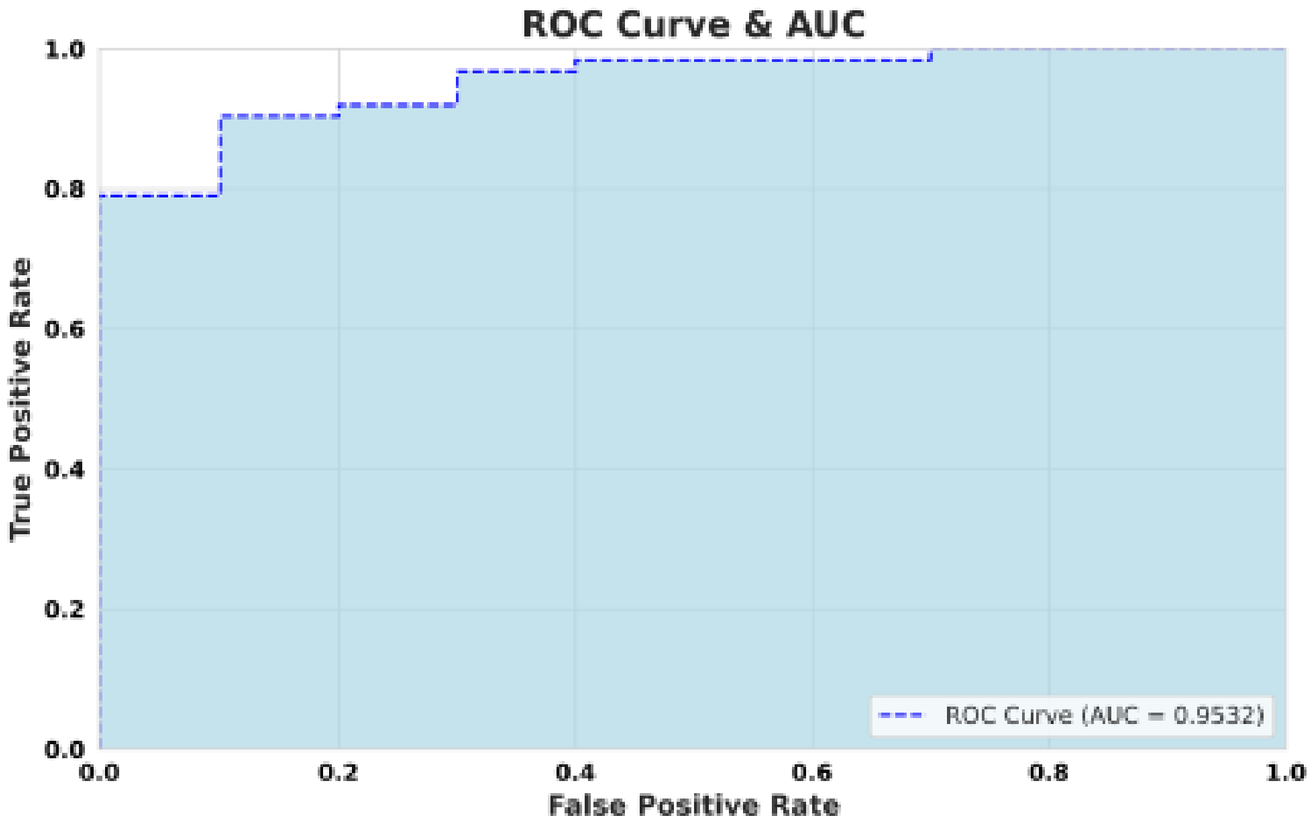
- Evaluated machine learning models: Logistic Regression, Decision Tree, Random Forest, SVM, Naive Bayes, and K-Nearest Neighbors (KNN).
- Key metrics used: Accuracy, Precision, Recall, and F1-score, suitable for binary classification.
- Logistic Regression and KNN achieved the highest accuracy of **93.06%** after hyperparameter tuning.
- Decision Tree showed lower accuracy at approximately 86.11%.



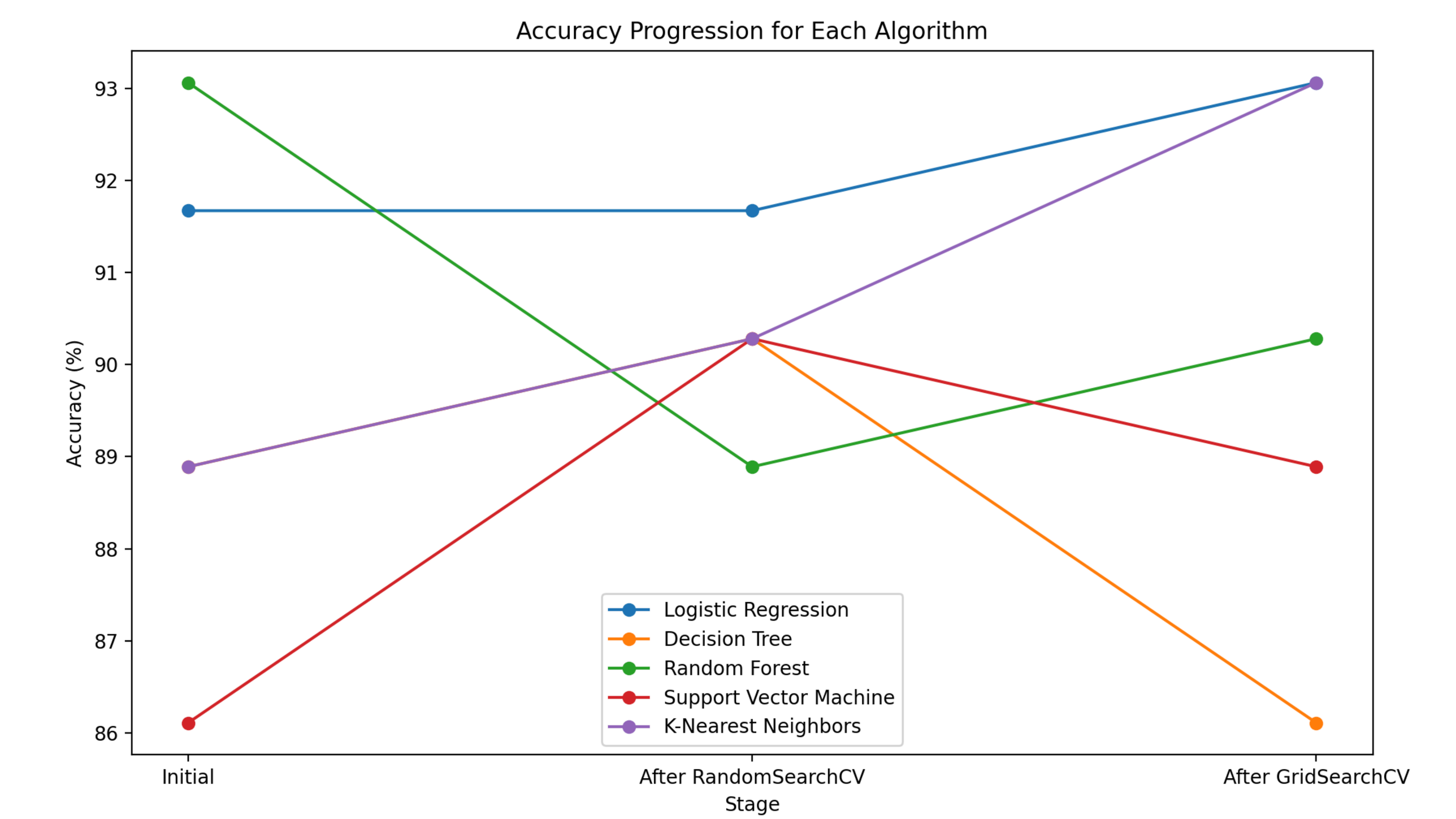
Performance Metrics & Results

COMPARISON OF CLASSIFICATION ALGORITHM ACCURACIES

Algorithm	Accuracy - Initial Model Classifier Performance	Accuracy - After RandomSearchCV	Accuracy - After GridSearch CV
Logistic Regression	91.67%	91.67%	93.06%
Decision Tree	88.89%	90.28%	86.11%
Random Forest	93.06%	88.89%	90.28%
Support Vector Machine	86.11%	90.28%	88.89%
K-Nearest Neighbors	88.89%	90.28%	93.06%



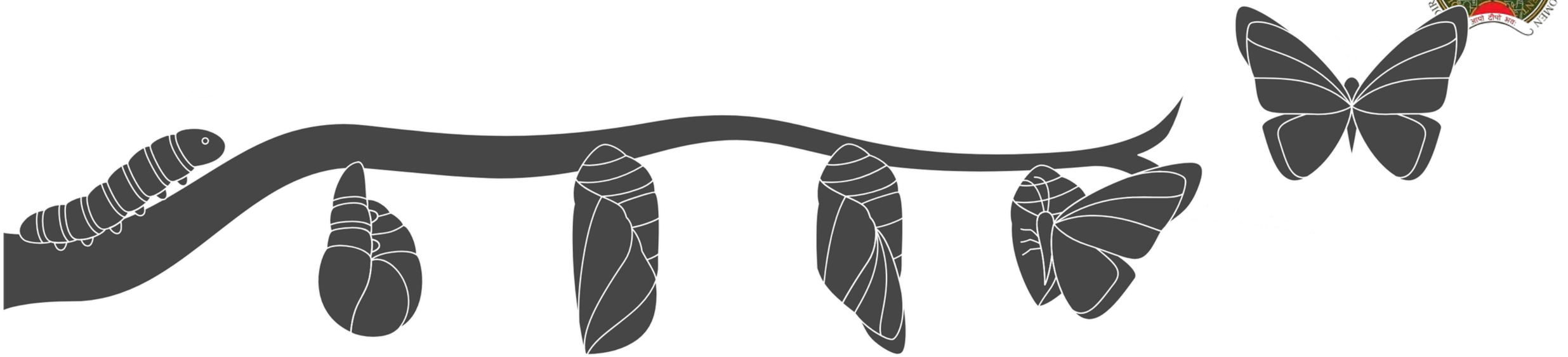
Analysis of the Top-Performing Model's ROC and AUC


Accuracy Progression for Each Algorithm
Accuracy (%)
93
92
91
90
89
88
87
86
Initial
After RandomSearchCV
After GridSearchCV
Stage
Logistic Regression
Decision Tree
Random Forest
Support Vector Machine
K-Nearest Neighbors

Interpretation of Results

- The Logistic Regression and K-Nearest Neighbors (KNN) models demonstrated the highest overall accuracy (~93%), indicating strong potential for lung cancer detection in the dataset used.
- KNN excelled with optimized parameters (weights='distance', n_neighbors=7, p=1), providing balanced precision and recall, notably detecting positives effectively.
- Random Forest and Support Vector Machine (SVM) showed sensitivity to class imbalance, leading to variability in recall values across classes.
- Decision Tree performed comparatively lower, possibly due to overfitting and less robustness to data variations.
- The AdaBoost algorithm helped identify critical features such as smoking status, age, chronic disease, fatigue, and yellow fingers, improving interpretability and stability of predictions.
- Despite strong quantitative performance, challenges persist in model interpretability and handling imbalanced data, common issues in medical datasets.
- The findings affirm the feasibility of AI-assisted lung cancer screening, encouraging further integration with larger, more diverse datasets and multi-modal data inputs for enhanced generalization and clinical adoption.

Future Work



CURRENT MODEL

Data Integration

Integrate multimodal data sources such as medical imaging (CT, PET), genomic data, and electronic health

Dataset Expansion

To enhance model generalization across populations differing in geography, ethnicity, and socioeconomic status

Privacy & Interpretability

Address privacy concerns with federated learning and develop more interpretable AI models for clinical adoption.

Continuous Improvement

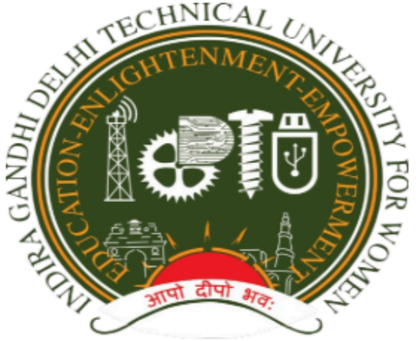
improvement in preprocessing, augmentation, and tuning to further boost prediction accuracy and reliability.

FUTURE READY

Accurate, reliable, and interpretable AI based diagnosis system

References

- World Health Organization, “Cancer Fact Sheet.”
- Bade, B. C., & Cruz, C. S. D. (2020). Lung cancer 2020: epidemiology, etiology, and prevention. Clinics in chest medicine, 41(1), 1.
- MacRosty, C. et al. (2020). Lung cancer in women: a modern epidemic. Clinics in chest medicine, 41(1), 53.
- Ahmad, A. S., & Mayya, A. M. (2020). A new tool to predict lung cancer based on risk factors. Heliyon.
- Patra, R. (2020). Prediction of Lung Cancer Using Machine Learning Classifier. In: Chaubey, N., Parikh, S., Amin, K. (eds) Computing Science, Communication and Security. COMS2 2020.
- Singh, G. et al. (2019). Performance analysis of ML approaches for detection/classification of lung cancer. Neural Computing and Applications, 31(10).
- Bartholomai, J., & Frieboes, H. (2018). Lung cancer survival prediction via ML regression, classification, and statistical techniques. IEEE ISSPIT.
- Faisal, M. et al. (2018). Evaluation of ML classifiers and ensembles in early lung cancer prediction. IEEE Conference.
- Dritsas, E., & Trigka, M. (2022). Lung Cancer Risk Prediction with Machine Learning Models. Big Data and Cognitive Computing.
- Maleki, N., & Niaki, S. T. A. (2023). Intelligent algorithm for lung cancer diagnosis using features from CT images.
- Islam, M. R. et al. (2022). Applying supervised contrastive learning for disease detection in medical imaging. Computers in Biology and Medicine.
- Alanazi, S., & Alanazi, R. (2025). Medical image classification with federated CNN and SMOTE. Alexandria Engineering Journal.
- Mir, A. et al. (2021). Hybrid quantum-classical models for medical image classification. Computers, Materials and Continua.
- Gencer, K. et al. (2025). Photodiagnosis with deep learning: GAN and autoencoder approach.
- Guefrachi, S. et al. (2025). Multistage CNN fine-tuning for medical classification. Arabian Journal for Science and Engineering.
- Huang, C. et al. (2024). Accurate detection using lightweight CNN and Fire Hawk Optimizer. Heliyon.
- LinkedIn advice on evaluating ML models: https://www.linkedin.com/advice/0/how-can-you-evaluate-machine-learning-models-using-iplye
- Dataset: https://www.kaggle.com/datasets/mysarahmadbhatt/lung-cancer
- Article: https://www.sciencedirect.com/science/article/pii/S2772442523000175
- Project GitHub: khushi-dutta/Lung-Cancer-Detection-Using-Machine-Learnin
- Gradio documentation: https://gradio.app



14th International Conference on
Computing, Communication and Sensor Network.CCSN2025

Thankyou

Presented by

Khushi Dutta
Manya Singh