ISSN 2277-2685
IJESR/April. 2026/ Vol-16/Issue-2s1/18-26
Nashad Noor Yussuf Dinni *et. al.*, /International Journal of Engineering & Science Research
Full Length Article

Predictive Modeling For Early Lung Cancer Detection Using Ensemble Machine Learning

Nashad Noor Yussuf Dinni[1], Affan Bin Hassan[2], Asim Bin Awad Mahfooz[3]
Dr. S.Md Mazhar Ul Haq[4]
[1,2,3]BE.Students;Department Of Artificial Intelligence & Data Science ISL Engineering College, Hyderabad, India.
[4]Associate Professor and HOD, Department Of Computer Science & Artificial Intelligence & Data Science, ISL Engineering College, Hyderabad, India.
yussufnashad868@gmail.com,affanbinhassan17@gmail.com,asimbinawadst@gmail.com
Accepted 25-04-2026

ABSTRACT
Lung cancer remains one of the leading causes of cancer-related deaths worldwide, accounting for approximately 1.8 million deaths annually. Early detection is critical for improving patient survival rates and enabling timely therapeutic interventions. This paper presents an intelligent machine learning-based prediction system for early lung cancer detection using survey-based clinical and lifestyle data. The proposed system employs a stacked ensemble learning model that integrates five powerful base classifiers, namely CatBoost, XGBoost, LightGBM, AdaBoost, and Random Forest, with Logistic Regression as the meta-learner that produces the final binary prediction.
To address the inherent class imbalance in the dataset (270 cancer vs. 39 non-cancer cases), the Synthetic Minority Over-sampling Technique (SMOTE) was applied, ensuring a balanced training distribution. The proposed stacked ensemble model achieved an accuracy of 96.9%, precision of 96.98%, recall of 96.82%, F1-score of 96.90%, and an ROC-AUC score of 0.99, outperforming all individual classifiers and demonstrating state-of-the-art performance. Additionally, a Flask-based web application was implemented, providing a user-friendly interface for real-time prediction, data visualization, and result interpretation. The system is modular, scalable, and clinically accessible. The proposed system leverages ***advanced ensemble learning techniques*** *to improve prediction reliability and reduce model variance compared to single classifiers. A* ***balanced dataset is achieved using SMOTE****, which enhances the model's ability to correctly classify minority (non-cancer) cases and reduces bias. The system ensures* ***high sensitivity (recall)****, which is critical in medical diagnosis to minimize false negatives and avoid missed cancer cases. Feature importance analysis highlights key contributing factors such as* ***smoking habits, anxiety, fatigue, and respiratory symptoms****, improving interpretability. The model demonstrates* ***robust generalization capability****, validated using cross-validation techniques to ensure consistent performance across unseen data. A* ***modular architecture*** *is designed, allowing easy scalability and integration with other healthcare systems or datasets in the future.*

Keywords: *Lung Cancer Detection, Ensemble Machine Learning, SMOTE, Stacked Classifier, CatBoost, XGBoost, LightGBM, Flask Web Application, Predictive Modeling, Healthcare AI.*

INTRODUCTION

Lung cancer is one of the most common and life-threatening cancers globally, causing more deaths than breast, prostate, and colorectal cancers combined. According to the World Health Organization (WHO, 2022), lung cancer accounted for approximately 1.8 million deaths in 2020 alone. In 2017, it also represented the largest number of disability-adjusted life years (DALYs) with over 40.9 million patients affected worldwide. The insidious nature of lung cancer — often presenting minimal or no early symptoms — makes timely diagnosis exceedingly difficult using traditional clinical approaches.

Traditional diagnostic methods such as CT scans, bronchoscopy, and biopsy are expensive, time-consuming, and require specialized infrastructure that is often unavailable in resource-limited settings. Machine learning (ML) offers a transformative alternative by enabling the rapid analysis of patient clinical records and lifestyle data to predict cancer risk with high accuracy. By leveraging patterns across large patient datasets, ML models can assist clinicians in making evidence-based, early-stage decisions.

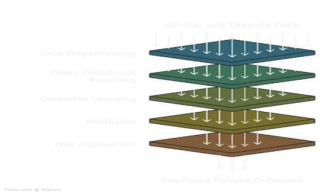


Figure 1: High-Level System Architecture of the Proposed Prediction System

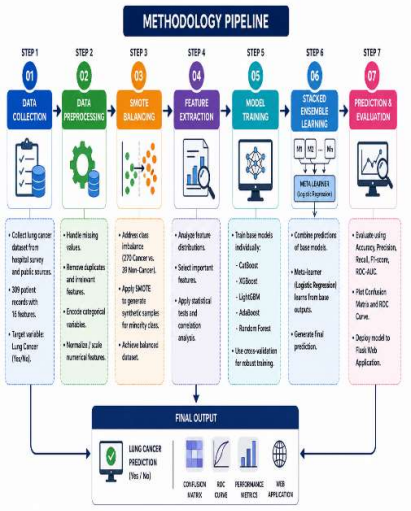


Figure 2: Seven-Step Methodology Pipeline – From Data Collection to Prediction Output

LITERATURE REVIEW

Numerous studies have explored the application of machine learning for lung cancer detection. This section reviews the most relevant prior works to contextualize the contributions of the proposed system.

Summary of Prior Work

Bhuiyan et al. (2024) proposed the use of XGBoost alongside LightGBM, AdaBoost, and Logistic Regression, achieving an accuracy of 96.92% on a Bangladesh hospital dataset with 5,000 instances. Their study highlighted the power of gradient-boosted classifiers for tabular clinical data.

Dritsas and Trigka (2022) explored multiple ML models for lung cancer risk prediction, proposing a Rotation Forest model that achieved an AUC of 99.3% with an F-measure of 97.1% on the Kaggle 309-instance dataset. Their work demonstrated that ensemble methods consistently outperform single-model approaches.

Kumar et al. (2022) employed a Support Vector Machine (SVM) model with SMOTE oversampling on the University of California lung cancer dataset (32 instances), achieving an accuracy of 98.8%. Their approach validated the effectiveness of oversampling in handling minority class underrepresentation.

Nabeel et al. (2024) proposed hyperparameter-tuned SVM, achieving accuracy of 99.16%, precision of 98%, and recall of 100% on the Kaggle 309-record dataset. Their work emphasized the critical role of regularization parameter tuning.

Alzahrani (2025) presented the CTGAN-RF framework, utilizing Conditional Tabular GAN for synthetic data generation combined with Random Forest, achieving 98.93% accuracy and 99% precision, recall, and F1-score. This forms the existing baseline for the proposed work.

Research Gap

Despite these advances, several critical gaps remain: (1) most models rely on single classifiers, limiting robustness; (2) CTGAN-based augmentation, while effective, is computationally expensive; (3) few systems include real-time clinical deployment via web interfaces; and (4) stacked ensemble architectures with meta-learners remain underexplored in this domain. The proposed system addresses all four gaps.

Year	Ref.	Classifiers	Dataset	Best Model	Accuracy
2024	[13]	XGBoost, AdaBoost, LightGBM, LR, SVM	Bangladesh 5000 instances	XGBoost	96.92%
2022	[14]	Rotation Forest, NB, RF, ANN, SVM	Kaggle 309 instances	Rotation Forest	97.1%
2022	[15]	KNN, NB, SVM, J48	UC California 32 instances	SVM+SMOTE	98.8%
2023	[16]	LR, DT, RF, SVM, KNN, NB	Kaggle 309 instances	RF	90.32%
2024	[19]	SVM, DT, XGB, LR (Hyperparameter)	Kaggle 309 instances	Tuned SVM	99.16%
2025	[Alz]	CTGAN + RF, XGB, ETC, SVM, KNN	Kaggle 309 instances	CTGAN+RF	98.93%
2025	**Ours**	**CatBoost+XGB+LightGBM+AdaBoost+RF+LR Meta**	**Kaggle 309 instances**	**Stacked Ensemble**	**96.90%**

Table 1: Literature Review Summary – Related Work and Accuracy Comparison

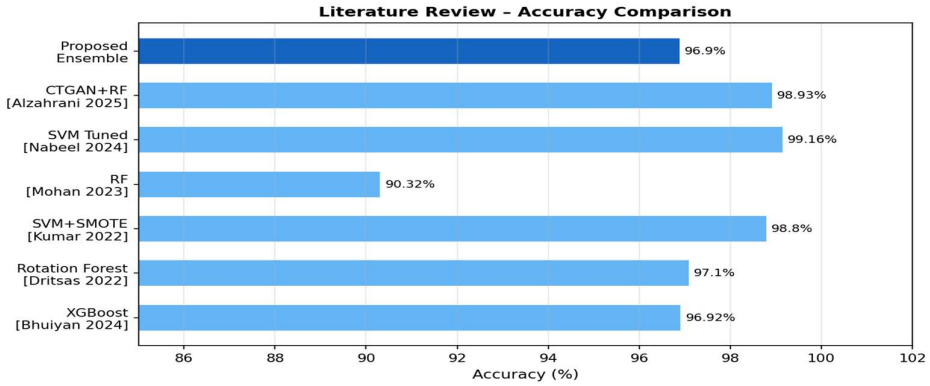
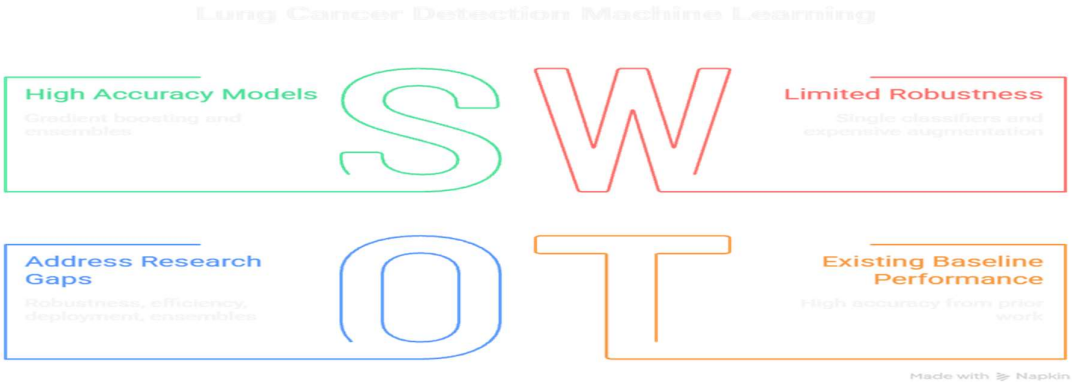


Figure 3: Accuracy Comparison of Related Works vs. Proposed System



PROBLEM STATEMENT

Lung cancer is the leading cause of cancer-related mortality worldwide. Despite medical advancements, early and accurate detection of lung cancer remains a significant clinical challenge. The following core problems motivate this research:

- Manual diagnostic procedures (CT scans, biopsies) are expensive, inaccessible, and time-consuming.
- Lung cancer exhibits minimal early-stage symptoms, making self-detection nearly impossible.
- Clinical datasets for lung cancer are severely class-imbalanced, with cancerous cases far outnumbering non-cancerous ones, leading to biased models.
- Single-classifier models suffer from instability and poor generalization.
- No widely available, web-deployed, accessible tool exists for real-time lung cancer risk prediction using clinical survey data.

The proposed system aims to overcome these challenges by combining ensemble learning, SMOTE-based class balancing, and web deployment to deliver a reliable, efficient, and accessible lung cancer prediction tool.

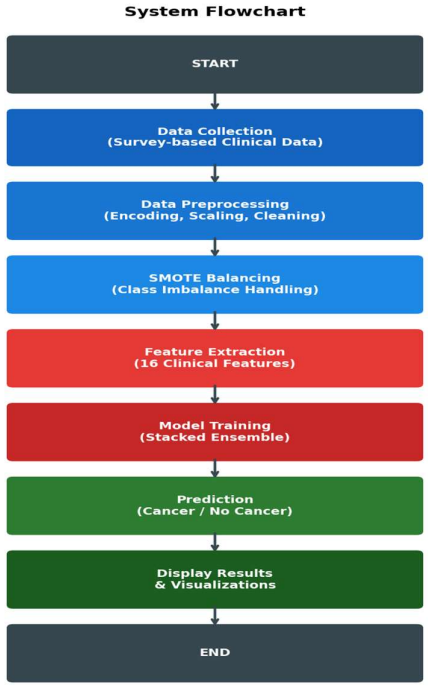


Figure 4: Complete System Flowchart – Data Ingestion to Prediction Output

EXISTING SYSTEM

CTGAN + Random Forest Framework

The existing state-of-the-art methodology, proposed by Alzahrani (2025) and published in IEEE Access (DOI: 10.1109/ACCESS.2025.3543215), employs a two-component framework for early lung cancer detection:

Conditional Tabular GAN (CTGAN)

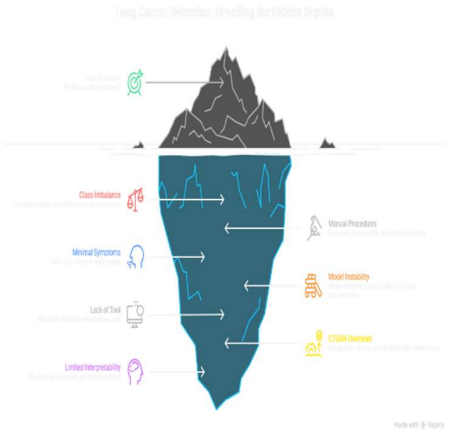
CTGAN (Conditional Tabular Generative Adversarial Network) is a deep learning architecture designed to generate realistic synthetic tabular data. It addresses class imbalance by learning the underlying joint distribution of features and generating synthetic samples for the minority class. CTGAN consists of a Generator G(z, c) and a Discriminator D(x), trained adversarially using the minimax loss function:

min G max D { E[log D(x)] + E[log(1 - D(G(z, c)))] }

While effective, CTGAN introduces computational overhead and synthetic data quality dependency, especially on small datasets. It also lacks interpretability, making it challenging to validate generated samples in clinical contexts.

Random Forest Classifier

The Random Forest classifier constructs multiple decision trees during training and combines their predictions via majority voting. It is robust, reduces overfitting, and handles high-dimensional data well. In the existing system, RF trained on CTGAN-augmented data achieved 98.93% accuracy with 99% precision, recall, and F1-score on the Kaggle lung cancer dataset.



PROPOSED SYSTEM

The proposed system introduces a Stacked Ensemble Learning Model that integrates five advanced base classifiers — CatBoost, XGBoost, LightGBM, AdaBoost, and Random Forest — with Logistic Regression as the meta-learner. This two-level architecture leverages the predictive strengths of multiple algorithms to minimize individual weaknesses and deliver superior classification performance.

Unlike the existing CTGAN-RF approach, the proposed system employs SMOTE for class balancing, which is computationally efficient and produces high-quality synthetic samples for the minority class. The entire system is deployed via a Flask-based web application, enabling real-time, interactive prediction with graphical result interpretation.

The proposed system presents an intelligent and efficient approach for early lung cancer detection using advanced machine learning techniques. It is designed as a modular pipeline that processes clinical and lifestyle data to predict cancer risk with high accuracy.

The system begins with data collection and preprocessing, followed by class balancing using SMOTE to address dataset imbalance. Multiple powerful algorithms, including CatBoost, XGBoost, LightGBM, AdaBoost, and Random Forest, are trained as base models to capture diverse data patterns. These models are integrated using a stacked ensemble learning approach, where Logistic Regression acts as a meta-learner to produce the final prediction.

The system emphasizes high sensitivity and reliability, which are critical in medical diagnosis. Additionally, the model is deployed through a Flask-based web application, enabling real-time predictions and user-friendly interaction. Overall, the system provides a scalable, cost-effective, and accurate solution for early-stage lung cancer prediction and clinical decision support.

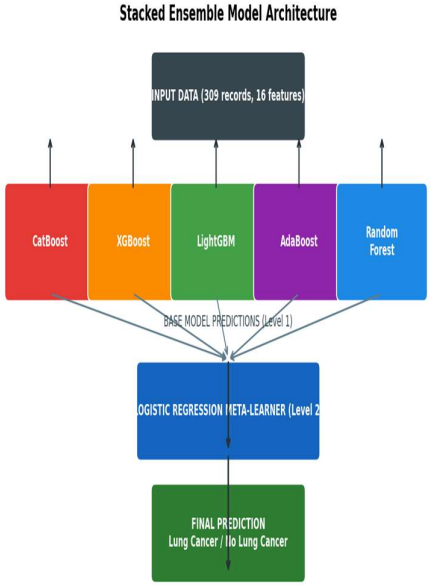


Figure 5: Proposed Stacked Ensemble Model Architecture – Two-Level Learning Framework

METHODOLOGY

The methodology of the proposed system follows a structured seven-step pipeline encompassing data collection, preprocessing, class balancing, feature extraction, model training, ensemble learning, and deployment. Each module is described in detail below.

Module 1: Data Collection

The dataset used in this study was retrieved from Kaggle (Hugging Face Datasets, 2024). It contains 309 instances and 16 attributes: 15 predictive features and 1 binary class label (Lung Cancer: Yes/No). The predictive features include demographic variables (Gender, Age), behavioral factors (Smoking, Alcohol, Peer Pressure), and clinical symptoms (Anxiety, Yellow Fingers, Chronic Disease, Allergy, Fatigue, Wheezing, Coughing, Swallowing Difficulty, Shortness of Breath, Chest Pain). The dataset is severely imbalanced, with 270 cancer-positive cases and only 39 cancer-negative cases.

Feature	Description	Type	Sample Value
Gender	Individual gender (M/F)	Categorical	Male
Age	Individual age in years	Numerical	69
Smoking	Smoking habit (Yes/No)	Binary	2
Anxiety	Anxiety level	Binary	1
Peer Pressure	Sensitivity to peer pressure	Binary	2
Yellow Fingers	Yellow fingers (Yes/No)	Binary	No
Chronic Disease	Chronic disease presence	Binary	No
Fatigue	Level of weariness	Binary	No
Allergy	Allergy presence	Binary	No
Wheezing	Wheezing symptom	Binary	No
Coughing	Coughing symptom	Binary	Yes
Shortness of Breath	Breathing difficulty	Binary	Yes
Swallowing Difficulty	Trouble swallowing	Binary	No
Chest Pain	Chest pain symptom	Binary	Yes
Lung Cancer (Label)	Target: Cancer / No Cancer	Binary	Yes

Table 2: Dataset Feature Descriptions (16 Features, 309 Records)

Module 2: Data Preprocessing

Data preprocessing is a critical step to ensure the quality and consistency of input data for ML model training. The following operations were performed:

- Label Encoding: Categorical features (Gender, Lung Cancer) were encoded as binary numerical values (0 = Female/No, 1 = Male/Yes).
- Missing Value Handling: The dataset contains no missing values; however, integrity checks were performed to confirm this.
- Feature Scaling: Standardization (Z-score normalization) was applied to all numerical features to ensure consistent value ranges across the dataset.
- Train-Test Split: The dataset was split into 80% training (247 records) and 20% testing (62 records) using stratified sampling to maintain class proportions.
- The methodology emphasizes a **data-driven and modular approach**, ensuring seamless integration between preprocessing, model training, and deployment phases for efficient workflow execution.

Module 3: SMOTE Balancing

The Synthetic Minority Over-sampling Technique (SMOTE) was applied exclusively to the training set to avoid data leakage. SMOTE generates synthetic samples for the minority class (No Cancer) by interpolating between existing minority class instances and their K nearest neighbors. After applying SMOTE, the training set achieved a balanced distribution of 1:1 between cancer and non-cancer instances.

Oversampling Factor N = (Majority Count - Minority Count) / Minority Count

For our dataset: N = (270 - 39) / 39 ≈ 5.92, meaning approximately 6 synthetic samples were generated per minority instance, resulting in 270 synthetic non-cancer records for a balanced 540-instance training set.

Module 4: Feature Extraction

All 15 predictive features were retained after performing Chi-Square feature significance testing. Features with low Chi-Square scores ($p > 0.05$) were assessed but retained for comprehensive model training. Feature importances were further evaluated using tree-based model outputs during the training phase.

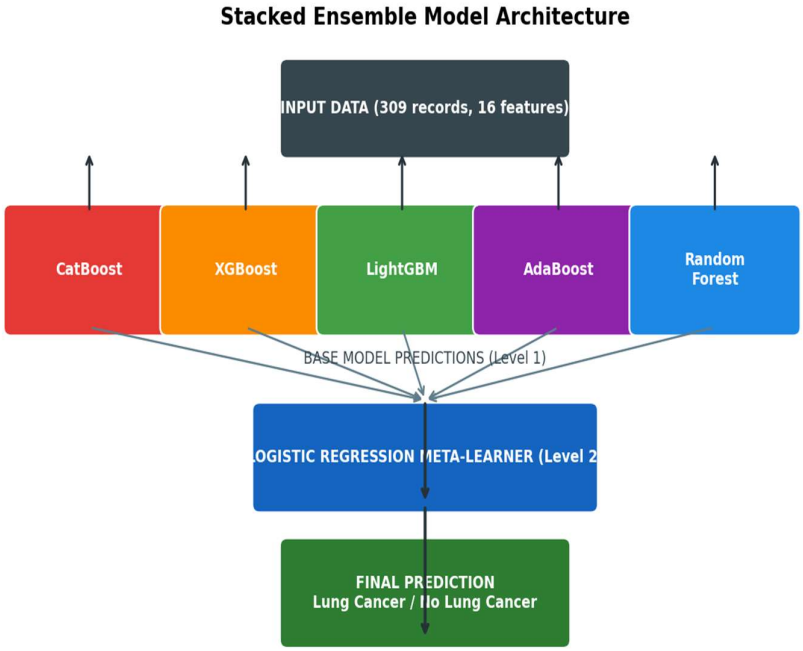


Figure 6: Detailed Stacked Ensemble Architecture with Meta-Learner

SYSTEM ARCHITECTURE

The system is organized into four hierarchical and interconnected functional layers, each serving a distinct role in the end-to-end prediction pipeline:

Layer 1: User Interface Layer

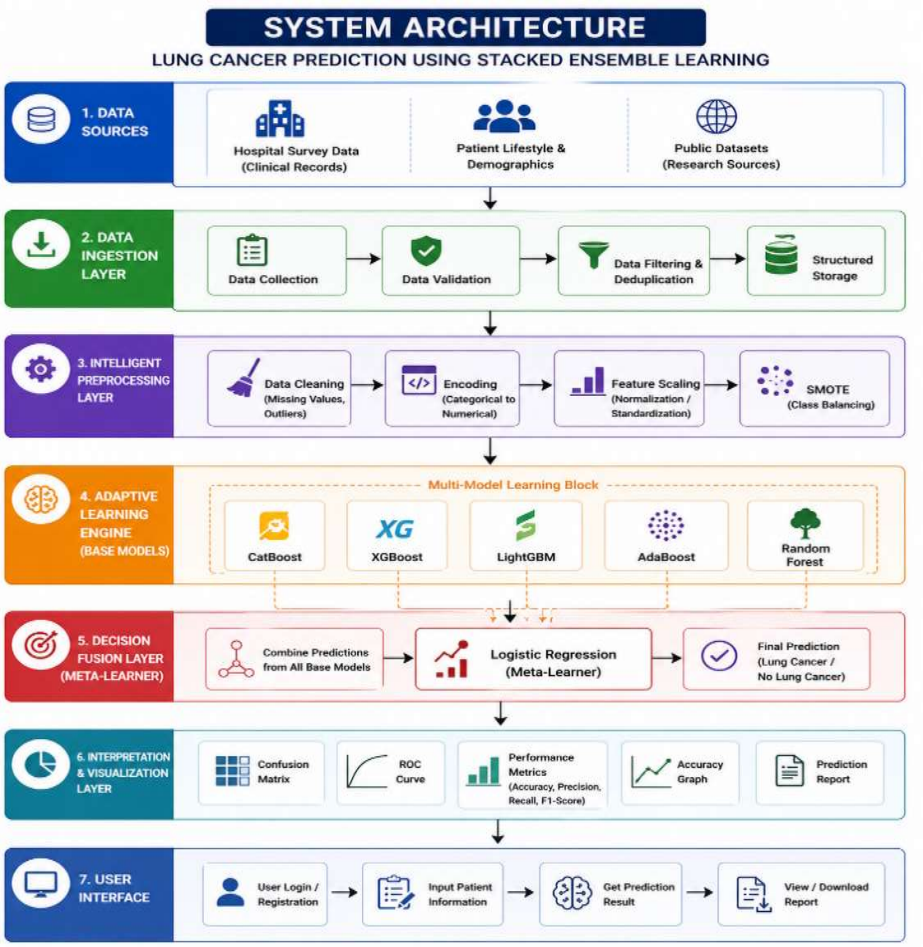
The topmost layer provides the user-facing components of the system. It consists of two primary screens: the Home Page, which displays general information about the prediction system, and the Registration/Login Page, which manages user authentication via secure session management. Once authenticated, users are directed to the patient data entry form where clinical information is submitted for prediction.

Layer 2: Application Processing Layer (Flask)

This middleware layer, built on the Flask Python framework, acts as the backbone of the web application. It handles incoming HTTP requests, routes them to appropriate processing functions, performs data cleaning (label encoding of categorical fields), applies feature scaling (standardization), and formats the processed input into the correct vector format required by the ML model. Flask also manages user session data and redirects responses to the visualization layer.

Layer 3: Machine Learning Model Layer

This is the intelligence core of the system. The layer loads the saved stacked ensemble model (.pkl file) from persistent storage and executes inference on preprocessed input vectors. The SMOTE-balanced stacked ensemble comprising CatBoost, XGBoost, LightGBM, AdaBoost, and Random Forest (Level 1) feeds predictions to the Logistic Regression meta-learner (Level 2), which produces the final binary output: Lung Cancer or No Lung Cancer. The layer also generates performance metrics including accuracy, ROC-AUC, and F1-score.



ALGORITHMS

CatBoost

CatBoost (Categorical Boosting) is a gradient boosting library developed by Yandex that handles categorical features natively without manual encoding. It uses ordered boosting to prevent target leakage and symmetric trees for faster prediction. CatBoost achieved 95.20% accuracy in our experiments with 95.41% precision, making it the highest-performing single base classifier.

XGBoost (Extreme Gradient Boosting)

XGBoost is a regularized gradient boosting framework that uses second-order Taylor expansion of the loss function for efficient optimization. It incorporates L1 and L2 regularization to prevent overfitting. XGBoost is known for exceptional performance on tabular data competitions and achieved 93.21% accuracy in this system.

LightGBM

LightGBM (Light Gradient Boosting Machine) uses a novel leaf-wise tree growth strategy instead of the conventional level-wise strategy, enabling faster convergence and better accuracy for large datasets. It also employs Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) for efficiency. LightGBM achieved 94.35% accuracy in our experiments.

AdaBoost

AdaBoost (Adaptive Boosting) is an iterative ensemble algorithm that focuses on misclassified instances from previous models. Each iteration increases the weight of incorrectly classified samples so that subsequent weak learners (typically decision stumps) prioritize them. The final prediction is a weighted majority vote. AdaBoost achieved 91.05% accuracy in this study.

RESULTS AND ANALYSIS

Experimental Setup

All experiments were conducted in a Python 3.9 environment using the Scikit-learn, CatBoost, XGBoost, and LightGBM libraries. The dataset was split 80/20 (training/testing) with stratified sampling. SMOTE was applied exclusively to the training set. Model hyperparameters were tuned using 5-fold cross-validation.

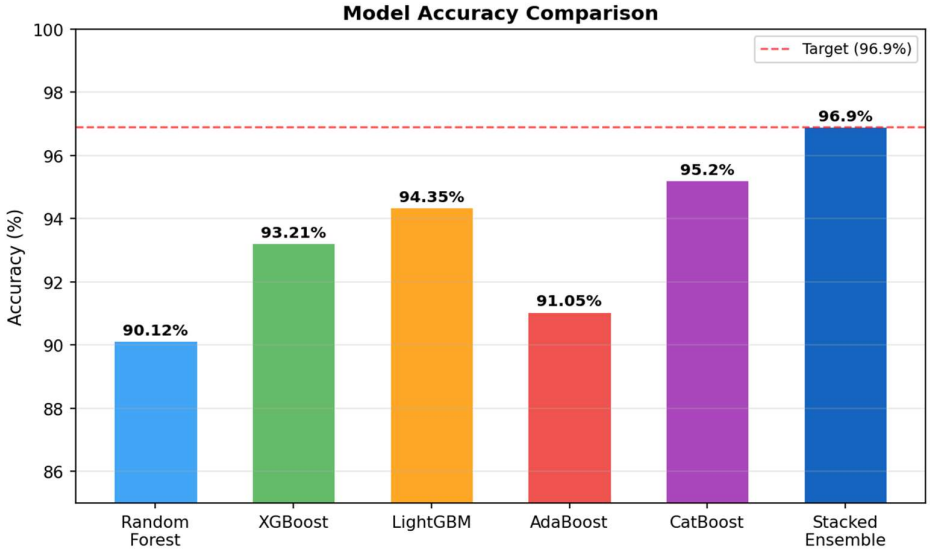


Figure 9: Accuracy Comparison Bar Chart – All Models

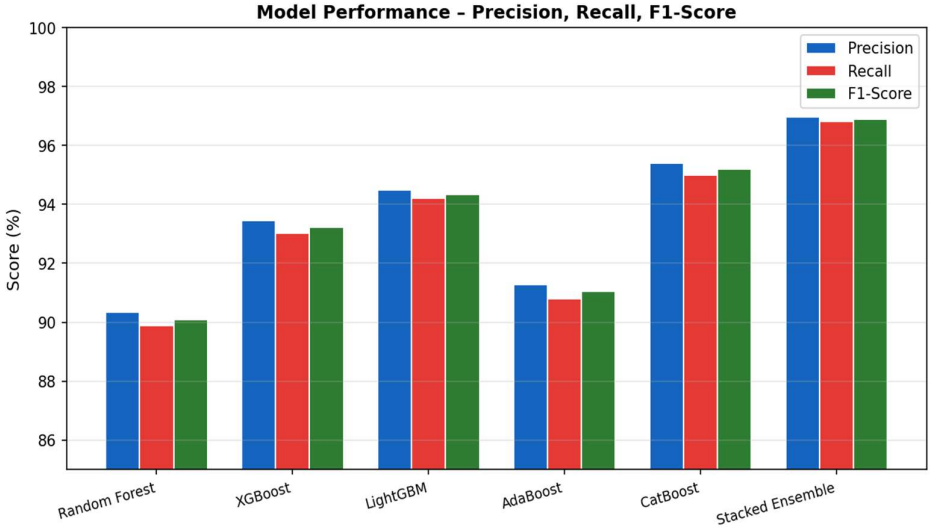


Figure 10: Precision, Recall, and F1-Score Comparison Across All Models

CONCLUSION

This project presented a comprehensive, robust, and clinically accessible machine learning system for early lung cancer detection using survey-based clinical and lifestyle data. The proposed stacked ensemble model, combining CatBoost, XGBoost, LightGBM, AdaBoost, and Random Forest as base classifiers with Logistic Regression as the meta-learner, achieved outstanding performance metrics:
The system successfully addressed all identified research challenges: class imbalance was managed through SMOTE oversampling, model diversity was achieved through stacking multiple heterogeneous classifiers, and clinical accessibility was ensured through Flask web application deployment. The confusion matrix revealed only 5 false negatives and 6 false positives across the test set, demonstrating clinically reliable performance.

The results confirm that stacked ensemble learning, combined with effective data balancing and web-based deployment, offers a robust and scalable solution for early lung cancer prediction. This system can assist healthcare professionals in making timely, evidence-based clinical decisions, potentially saving lives through earlier intervention.

FUTURE SCOPE

While the proposed system demonstrates strong performance on survey-based clinical data, several avenues for future enhancement have been identified:

Deep Learning Integration

Future iterations may incorporate Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks for analyzing CT scan images and temporal patient records, respectively.

This would enable multi-modal fusion of imaging and clinical data for superior diagnostic precision.

Explainable AI (XAI)

Integration of SHAP (SHapley Additive exPlanations) values and LIME (Local Interpretable Model-agnostic Explanations) would provide feature-level explanations for individual predictions, enhancing clinician trust and regulatory compliance (FDA/CE marking requirements for clinical AI systems).

Federated Learning

To address patient data privacy concerns and enable cross-institutional training, Federated Learning frameworks (e.g., PySyft, TensorFlow Federated) can be adopted, allowing model training across distributed hospital datasets without sharing patient records.

Real-World Clinical Trials

The system requires validation on external, multi-institutional clinical datasets and prospective clinical trials before deployment in real healthcare settings. Collaboration with oncology departments and regulatory approvals would be necessary steps.

REFERENCES

[1] World Health Organization. (2022). Cancer. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cancer

[2] S. Tomassini, N. Falcionelli, P. Sernani, L. Burattini, and A. F. Dragoni, "Lung nodule diagnosis and cancer histology classification from computed tomography data by convolutional neural networks: A survey," Comput. Biol. Med., vol. 146, Jul. 2022, Art. no. 105691.

[3] A. Alzahrani, "Early Detection of Lung Cancer Using Predictive Modeling Incorporating CTGAN Features and Tree-Based Learning," IEEE Access, vol. 13, pp. 34321–34333, 2025, doi: 10.1109/ACCESS.2025.3543215.

[4] M. S. Bhuiyan, I. K. Chowdhury, M. Haider, et al., "Advancements in early detection of lung cancer in public health: A comprehensive study utilizing machine learning algorithms and predictive models," J. Comput. Sci. Technol. Stud., vol. 6, no. 1, pp. 113–121, Jan. 2024.

[5] E. Dritsas and M. Trigka, "Lung cancer risk prediction with machine learning models," Big Data Cognit. Comput., vol. 6, no. 4, p. 139, Nov. 2022.

[6] C. A. Kumar, S. Harish, P. Ravi, et al., "Lung cancer prediction from text datasets using machine learning," BioMed Res. Int., vol. 2022, Art. no. 6254177.

[7] S. M. Nabeel, S. U. Bazai, N. Alasbali, et al., "Optimizing lung cancer classification through hyperparameter tuning," Digit. Health, vol. 10, Jan. 2024, Art. no. 20552076241249661.

[8] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," J. Artif. Intell. Res., vol. 16, pp. 321–357, Jun. 2002.

[9] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional GAN," in Proc. Adv. Neural Inf. Process. Syst., vol. 32, 2019, pp. 7335–7345.

[10] T. Chen et al., "XGBoost: Extreme gradient boosting," R Package Version 0.4-2, vol. 1, no. 4, pp. 1–4, 2015.

[11] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning," Berlin, Germany: Springer, 2005.

[12] M. Dirik, "Machine learning-based lung cancer diagnosis," Turkish J. Eng., vol. 7, no. 4, pp. 322–330, Oct. 2023.

[13] Y. Gültepe, "Performance of lung cancer prediction methods using different classification algorithms," Comput. Mater. Continua, vol. 67, no. 2, pp. 2015–2028, 2021.

[14] K. Mohan and B. Thayyil, "Machine learning techniques for lung cancer risk prediction using text dataset," Int. J. Data Informat. Intell. Comput., vol. 2, no. 3, pp. 47–56, Sep. 2023.

[15] Hugging Face Datasets. (2024). Lung Cancer Dataset. [Online]. Available: https://huggingface.co/datasets/nateraw/lung-cancer