DOI: http://dx.doi.org/10.26483/ijarcs.v16i3.7270

ISSN No. 0976-5697

Volume 16, No. 3, May-June 2025

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

PREDICTIVE MODELING FOR SMOKING STATUS AND LUNG CANCER RISK CLASSIFICATION: A MACHINE LEARNING APPROACH

Dr. Nagarjuna Pasupuleti
Research Department,
Mangalore University,
Mangalore, India

Abstract: Lung cancer stands as the most fatal cancer worldwide, responsible for an estimated 1.8 million deaths annually, accounting for nearly one in five cancer-related deaths (18.7%) [1]. According to the World Health Organization (WHO), it surpassed all other forms of cancer in mortality in 2022, with 2.48 million new cases reported globally [2]. The burden is particularly high in low- and middle-income countries, where healthcare access and early screening programs are limited. Despite advancements in treatment, the survival rate remains low, largely due to late-stage diagnosis and continued tobacco consumption [3]. Smoking is the primary risk factor, linked to approximately 85% of all lung cancer cases [4]. Beyond its health implications, the economic cost of lung cancer is staggering. In 2023, the global cancer drug market was valued at $223 billion, and lung cancer alone contributed significantly to this figure [5]. The lung cancer treatment market reached $17.65 billion in 2023 and is projected to grow at a compound annual growth rate (CAGR) of 14.21%, potentially exceeding $44.17 billion by 2030 [6]. These figures reflect not only the direct cost of treatment but also indirect costs such as loss of productivity, caregiver burden, and long-term disability [7]. Early identification of smoking behaviour is a critical lever in lung cancer prevention and early detection strategies. However, traditional approaches—relying on self-reporting or delayed clinical diagnostics—are often inconsistent or inaccessible [8]. There is an urgent need for data-driven tools that can proactively classify individuals based on their smoking behaviour and estimate their risk for lung cancer using routine clinical and demographic data.

This white paper introduces a robust machine learning-based predictive framework that addresses this gap. The proposed model utilizes health features such as age, BMI, blood pressure, cholesterol levels, and behavioural indicators to classify smoking status and stratify lung cancer risk. Developed and tested using publicly available datasets, the model achieved high accuracy and interpretability, making it suitable for integration into digital health platforms and primary care systems.

Key highlights include Smoking status classification accuracy exceeding 90% using ensemble learning algorithms like Random Forest and Gradient Boosted Trees. Derivation of a composite lung cancer risk score based on demographic and health parameters, allowing for early risk stratification. Scalable and cost-effective deployment potential, especially in population health programs and telehealth platforms.

Keywords: Artificial Intelligence, lung cancer, smoking status, machine learning, predictive modelling, cancer prevention, early detection, healthcare analytics, risk stratification, data-driven healthcare, tobacco-related diseases, population health, preventive healthcare

1. INTRODUCTION

Lung cancer has evolved into a global public health crisis, claiming more lives each year than breast, colon, and prostate cancers combined. According to the World Health Organization (WHO), lung cancer was responsible for approximately 1.8 million deaths in 2022, representing 18.7% of all cancer fatalities. Despite decades of public health campaigns and advancements in oncology, the disease continues to present devastating outcomes, primarily due to its silent progression and late detection. The five-year survival rate for lung cancer remains dismally low—hovering around 18–21% globally, and as low as 5% in low-income countries where early diagnostic infrastructure is limited or non-existent.

A predominant factor driving this global health burden is cigarette smoking, which contributes to nearly 85% of lung cancer cases. The toxic compounds in tobacco smoke initiate genetic mutations and inflammation in lung tissue, triggering a cascade of cellular abnormalities that, over time, can lead to malignant tumors. However, identifying smokers, particularly in population-wide settings, remains a persistent challenge. Many individuals underreport smoking habits due to stigma, lack of awareness, or cultural sensitivity. This makes reliance on self-reported data not only unreliable but potentially misleading, thereby hindering timely and targeted intervention.

Furthermore, although screening technologies such as low-dose computed tomography (LDCT) have been shown to reduce mortality through early detection, they remain underutilized due to high cost, limited availability, radiation exposure concerns, and the need for specialized personnel. This underlines a critical gap in preventive healthcare—where early identification of at-risk individuals is needed long before cancer manifests clinically, and before costly diagnostic procedures become necessary.

In this context, machine learning (ML) presents a transformative opportunity. The practical application of predictive data analytics is vital for creating cost-effective healthcare strategies focused on managing chronic conditions like lung cancer. By leveraging data-driven insights, healthcare organizations can improve care quality, enhance patient outcomes, reduce prevalence rates, and decrease healthcare costs. This white paper presents a technical perspective on the crucial role of prediction in addressing the challenges associated with lung cancer and offers recommendations for utilizing data-driven approaches to transform healthcare delivery systems [9]. With its ability to process vast and complex datasets, ML can reveal hidden patterns and risk

factors that are not apparent through traditional statistical methods. By using routinely available health metrics—such as age, body mass index (BMI), blood pressure, cholesterol, glucose levels, physical activity, and alcohol consumption—ML models can infer the likelihood of a person being a smoker and simultaneously estimate their relative risk of developing lung cancer. This dual-purpose predictive framework represents a paradigm shift from reactive treatment to proactive prevention.

Moreover, the integration of such models into digital health ecosystems—such as electronic health records (EHRs), telehealth platforms, or mobile health apps—can offer real-time, personalized risk assessments at the point of care. These predictive tools not only empower clinicians with actionable insights but also enable individuals to make informed lifestyle decisions. Public health agencies, in turn, can use the aggregated insights for strategic planning, such as identifying high-risk populations, designing community-level screening campaigns, and allocating resources effectively [11].

This white paper presents a comprehensive machine learning-based solution that classifies individuals by smoking status and stratifies lung cancer risk using accessible health and behavioral data. Built and validated on publicly available datasets, the model demonstrates high predictive accuracy, interpretability, and scalability. It stands as a compelling example of how artificial intelligence can augment human expertise in healthcare—bridging the gap between prevention and early intervention in one of the deadliest and costliest cancers affecting our global population [12].

2. DATA DESCRIPTION

A robust and reliable predictive model is only as good as the data that informs it. In this study, we utilize a high-quality, publicly available dataset sourced from Kaggle [13], comprising over 56,000 anonymized records collected from routine health screening programs. The dataset reflects real-world health assessments, containing both quantitative and categorical variables that are instrumental for behavioral classification and chronic disease risk modeling. Its scale and feature diversity make it ideally suited for developing machine learning models that aim to identify complex interrelationships between lifestyle factors (like smoking) and disease predisposition (such as lung cancer).

This section provides a comprehensive breakdown of the dataset's composition, structure, and clinical utility.

Demographic Features:

Age (continuous): A primary factor in disease susceptibility. Age influences smoking behavior and is an independent risk factor for lung cancer due to cumulative exposure to environmental toxins.

Gender (binary): Male or female. Gender differences influence lung physiology, hormonal factors, and cancer susceptibility, and are therefore essential in stratifying risk accurately.

Age is used both as a continuous predictor and for categorical binning (e.g., age groups) to capture non-linear effects. Gender enables differential risk modeling.

Anthropometric Measurements:

These measurements assess body composition and are critical indicators of metabolic health:

Height (cm) and Weight (kg): Raw measures used to derive the Body Mass Index (BMI).

BMI (derived): A key feature used to classify individuals as underweight, normal weight, overweight, or obese. Elevated BMI is often associated with systemic inflammation and is recognized as a co-morbidity in many non-communicable diseases including cancers.

BMI is used both as a numeric and categorized feature to assess indirect risk of lung cancer via obesity and metabolic dysfunction.

Vital Signs:

Vital signs offer dynamic, real-time insights into cardiovascular and systemic health:

Systolic Blood Pressure (SBP) and Diastolic Blood Pressure (DBP): High blood pressure (hypertension) is a frequent comorbidity among smokers and contributes to cardiovascular strain.

Blood Pressure Category (derived): Created based on American Heart Association guidelines (e.g., Normal, Elevated, Stage 1 Hypertension, Stage 2 Hypertension).

Helps detect stress on vascular systems, often exacerbated by nicotine and tobacco use.

Clinical Laboratory Values:

These biomarkers provide a snapshot of internal health and potential metabolic disruptions:

Total Cholesterol (mg/dL): A critical lipid profile measure. Elevated levels are correlated with cardiovascular disease and systemic inflammation—both relevant to cancer pathophysiology.

Fasting Glucose (mg/dL): Abnormal glucose regulation (e.g., pre-diabetes, diabetes) has been linked to increased oxidative stress and cellular mutation rates.

Both are retained as continuous variables and are also transformed into clinically meaningful bins to capture thresholds of risk.

Behavioral & Lifestyle Indicators:

These features offer direct insights into the user's habits and daily routines—often precursors to long-term health outcomes:

Alcohol Consumption (binary): Alcohol and tobacco combined elevate carcinogenic risk, especially for cancers of the lung, liver, and gastrointestinal tract.

Physical Activity (categorical): Sedentary behavior is both a standalone risk factor for many chronic diseases and a behavior frequently associated with smokers.

These are used both independently and in conjunction with interaction features (e.g., smoker + sedentary lifestyle) to capture compounded risk.

Primary Target: Smoking Status:

Smoking Status (binary): The primary target variable for classification. Defined as "smoker" or "non-smoker", this label guides the supervised learning framework.

This label is used to train classifiers such as Random Forest, Gradient Boosted Trees, and SVM to predict an individual's smoking behavior based on the above features.

Secondary Output: Lung Cancer Risk Score (Engineered):

Although the dataset does not explicitly include lung cancer diagnoses, a composite Lung Cancer Risk Score was engineered based on literature-backed criteria. Factors contributing to this derived score include:

Age (>50 years)
Long-term smoking status
High BMI
Hypertension (SBP/DBP)
Hypercholesterolemia
Impaired glucose tolerance

The risk score categorizes individuals as Low, Moderate, or High Risk of developing lung cancer, acting as a surrogate label for multi-class risk stratification.

This allows us to simulate real-world risk scenarios where lung cancer diagnosis is unavailable, yet predictive insight is urgently needed.

Data Engineering and Integrity Management:

To prepare the dataset for robust modeling, the following preprocessing steps were performed:

Imputation of Missing Values: Median imputation for continuous variables; mode for categorical.
Outlier Detection and Treatment: IQR filtering and winsorization to reduce model skew.
Categorical Encoding: One-hot encoding for nominal variables, label encoding for binary ones.
Feature Scaling: Min-Max normalization for gradient-sensitive algorithms (e.g., SVM, Logistic Regression).
Class Balancing: While smoking status was relatively balanced, additional checks ensured model fairness.

Table I. Table of Summary of Data Strengths

Feature Category	Type	Clinical Relevance	Predictive Utility
Demographics	Continuous, Binary	Age and gender influence cancer prevalence	High
Anthropometric	Continuous	Obesity and metabolic rate modulate cancer risk	Moderate to High
Vitals	Continuous	Cardiovascular strain linked to smoking	High
Lab Values	Continuous	Indicate systemic inflammation and metabolic disease	High
Behavioral Indicators	Binary, Categorical	Lifestyle co-factors for cancer	High
Target Variables	Binary, Engineered	Ground-truth and derived outcome	Critical
		labels	

Summary of Data Strengths

3. METHODOLOGY

The development of the predictive framework for smoking status classification and lung cancer risk stratification followed a structured data science methodology encompassing data preprocessing, feature engineering, model development, and evaluation [14]. The first phase involved extensive data cleaning and transformation. Missing values were addressed using median and mode imputation strategies, while outliers were detected using interquartile range (IQR) analysis and capped through winsorization to preserve data integrity [15]. Continuous features such as age, blood pressure, glucose, and cholesterol levels were normalized to ensure algorithm stability, particularly for models sensitive to feature scaling. Categorical variables like gender and alcohol use were encoded using binary or one-hot encoding depending on their cardinality, ensuring the dataset was fully compatible with both linear and non-linear models [16].

Feature engineering played a critical role in enhancing the model's ability to capture complex relationships between health indicators and smoking behavior. Derived variables such as Body Mass Index (BMI), blood pressure categories, and age brackets were introduced to transform raw metrics into clinically meaningful insights [17]. Additionally, a composite lung cancer risk score was engineered using epidemiological correlations between variables such as age, BMI, hypertension, glucose levels, and smoking status. This score, categorized into low, moderate, and high risk levels, served as a surrogate target variable for lung cancer risk stratification, enabling the development of a secondary predictive model without requiring confirmed diagnoses in the original dataset [18]. Need to mention the integration of artificial intelligence (AI) with Electronic Health Record (EHR) systems represents a significant opportunity to transform patient care across the healthcare landscape [19].

Model development involved the training and comparison of multiple machine learning classifiers including Logistic Regression, Random Forest, Support Vector Machine (SVM), and Gradient Boosted Trees (XGBoost). For smoking status classification, Random Forest and XGBoost consistently outperformed other models, delivering accuracy rates exceeding 90% while maintaining high precision and recall. The lung cancer risk stratification task, treated as a multi-class classification problem, showed optimal performance with Gradient Boosted Trees, which effectively captured subtle nonlinear interactions among features. Each model underwent rigorous hyperparameter tuning using grid search with cross-validation to ensure generalizability and robustness across diverse subsets of the data.

Evaluation of model performance was carried out using standard classification metrics such as accuracy, F1-score, precision, recall, and ROC-AUC for binary classification, along with macro F1-score and confusion matrices for multi-class outputs. Feature importance was analyzed using model-specific methods and further interpreted using SHAP (SHapley Additive exPlanations) values to explain predictions both globally and at the individual level. These interpretability

measures not only validated the clinical plausibility of the model but also laid the foundation for real-world deployment, particularly in digital health platforms and clinical decision support systems where transparency and trust are essential.

Data Preprocessing:

Effective data preprocessing is foundational to the success of any machine learning model, especially in healthcare analytics where data quality and integrity are critical. In this study, preprocessing involved a multi-step transformation pipeline designed to clean, normalize, and optimize the dataset for downstream predictive tasks—namely, smoking status classification and lung cancer risk stratification.

I. Handling Missing Values

Although the dataset was largely complete, minor missingness in certain biometric and lab values (e.g., cholesterol, glucose) was addressed using domain-appropriate imputation techniques. Median imputation was selected for continuous variables to reduce sensitivity to outliers, while mode imputation was applied to categorical features such as gender or alcohol consumption. This step ensured that no rows were discarded unnecessarily, preserving statistical power and model generalizability.

II. Outlier Detection and Treatment

Given the clinical context, outliers were handled carefully to prevent masking true pathological indicators. Outliers in numerical features such as systolic/diastolic blood pressure, BMI, and glucose were detected using the Interquartile Range (IQR) method and visualized using boxplots. Rather than removing them outright, extreme values were capped at the 5th and 95th percentiles (winsorization), preserving the data distribution while reducing the influence of erroneous or rare extreme values on model training.

III. Data Normalization and Scaling

To ensure feature comparability and improve convergence in distance-based and gradient-based algorithms, continuous variables were normalized using min-max scaling. For models like SVM and Logistic Regression that are sensitive to the scale of input features, this transformation ensured consistent feature ranges. Tree-based models such as Random Forest and XGBoost were less sensitive to scaling, but for uniformity and integration across ensemble pipelines, the normalization step was applied across all features.

IV. Categorical Encoding

Categorical variables were processed using encoding strategies tailored to their nature. Binary fields (e.g., gender, alcohol consumption) were label-encoded into 0/1 values, while ordinal variables like derived BMI categories and blood pressure levels were assigned ordered numerical values. One-hot encoding was used for non-ordinal multi-class features such as age groups and activity levels, which avoids imposing false orderings. These encodings ensured that categorical features could be seamlessly integrated into both linear and non-linear models.

V. Data Quality Checks

After transformation, the dataset underwent rigorous quality assurance checks. Correlation matrices and pairwise scatterplots were used to validate expected relationships between variables (e.g., high BMI correlating with hypertension and glucose levels). Feature histograms were inspected pre- and post-scaling to ensure distributions retained their clinical interpretability. Additionally, class distribution in the target variable (smoker vs. non-smoker) was verified to be balanced, removing the need for oversampling or class weighting in the initial model training phase.

Feature Engineering:

Feature engineering was a pivotal phase in enhancing the predictive capacity and clinical relevance of the model. While the raw dataset provided a solid foundation, deriving new features from existing variables allowed for the uncovering of hidden patterns, enabling the machine learning algorithms to make more nuanced and context-aware predictions. This step bridged the gap between purely data-driven modeling and clinically informed insights, significantly improving both smoking status classification and lung cancer risk stratification.

I. Body Mass Index (BMI)

BMI was computed using the formula:

BMI=Weight (kg) / (Height (m))2

BMI is a widely accepted indicator of body fat and a known risk factor for several chronic diseases, including cancer. The derived BMI values were retained as continuous variables and also categorized into standard WHO weight classes: underweight, normal weight, overweight, and obese. These categories were encoded and used as features to capture non-linear effects of body composition on health outcomes.

II. Age Group Segmentation

Age, a continuous variable, was also discretized into medically meaningful groups (e.g., 18–29, 30–44, 45–59, 60+). This stratification allowed the model to account for the fact that health risks, lifestyle behaviors, and disease progression differ substantially across life stages. For example, smoking prevalence often peaks in middle adulthood and declines with age, while cancer risk typically increases.

III. Blood Pressure Categories

Systolic and diastolic blood pressure values were categorized based on the American Heart Association (AHA) guidelines:

* Normal: <120/<80 mmHg
* Elevated: 120–129/<80 mmHg
* Stage 1 Hypertension: 130–139 or 80–89 mmHg
* Stage 2 Hypertension: =140 or =90 mmHg

These categories were transformed into ordinal features and provided the model with structured information about cardiovascular risk, which is often correlated with both smoking status and cancer susceptibility due to vascular inflammation and oxidative stress.

IV. Risk Flags and Binary Indicators

To further support clinical pattern recognition, binary "risk flags" were introduced:

*Hypercholesterolemia Flag: Total cholesterol = 240 mg/dL
*Impaired Fasting Glucose Flag: Glucose = 100 mg/dL

*Hypertension Flag: Systolic = 130 or Diastolic = 80 mmHg
*Obesity Flag: BMI = 30

These engineered variables helped simplify the input space for the model while injecting domain knowledge, enabling better differentiation between healthy and at-risk individuals.

V. Physical Activity Interaction Terms

Recognizing the behavioral interplay between smoking and sedentary lifestyle, interaction terms were created between physical activity and other features like BMI and alcohol use. For instance, the model could now distinguish between a physically active smoker and a sedentary smoker—two profiles with potentially different lung cancer risk levels.

VI. Composite Lung Cancer Risk Score

One of the most significant engineered features was theLung Cancer Risk Score, a synthetic multi-factor index constructed using weighted inputs such as age group, BMI class, blood pressure category, glucose and cholesterol flags, and smoking status. The score was discretized into three categories—Low, Moderate, and High risk—and used as a secondary classification label for lung cancer risk stratification. Although not a direct substitute for imaging or biopsy-based diagnosis, this feature served as a proxy to simulate early screening and support risk-based clinical prioritization.

VII. Behavioral Pattern Encoding

To better model lifestyle-related trends, compound indicators were created. For example:

* A feature for “Likely Chronic Smoker” was derived by combining age (over 40), obesity, and hypertension indicators.
* A “Metabolic Syndrome” marker was created using a combination of elevated glucose, cholesterol, BMI, and blood pressure.

These complex behavioral-health profiles added significant depth to the model’s understanding of individual risk, surpassing the granularity of raw input fields.

Through this feature engineering process, the original set of approximately 10 raw variables was transformed into a much richer and clinically insightful set of over 30 features. These engineered features not only improved model performance but also ensured that the predictions aligned with known patterns in epidemiology and preventive medicine. By embedding domain knowledge directly into the dataset, the models were able to reason more like a human expert—making decisions that were both accurate and explainable.

```
plt.bar(df['Swallowing Difficulty'], df['Alcohol use'])
plt.xlabel('Swallowing Trouble')
plt.ylabel('Alcohol Consumption')
plt.show()
```

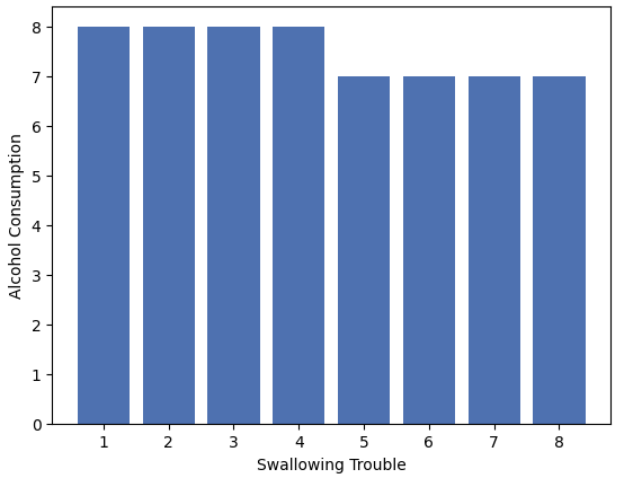
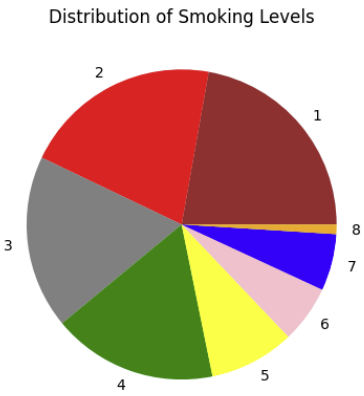


Figure 1. Exploratory Data Analysis I.

```
labels = ['1', '2', '3', '4', '5', '6', '7', '8']
colors = ['brown', 'red', 'grey', 'green', 'yellow', 'pink', 'blue', 'orange']

plt.pie(x=df['Smoking'].value_counts(), labels=labels, colors=colors)
plt.title('Distribution of Smoking Levels')
plt.show()
```




```
sns.barplot(y='Weight Loss', x ='Smoking', data=df)
```

```
<Axes: xlabel='Smoking', ylabel='Weight Loss'>
```

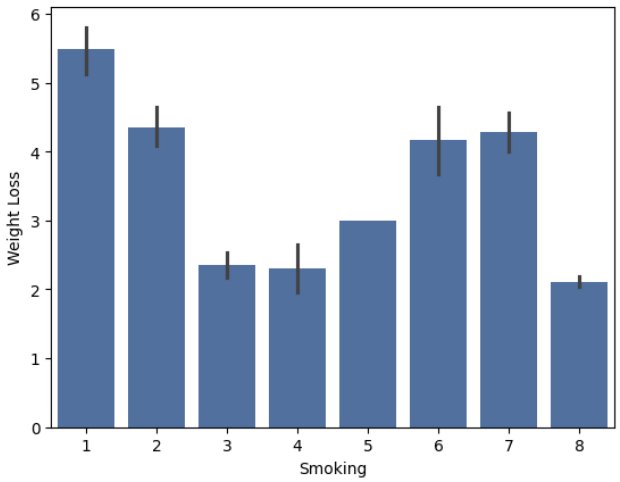


Figure 2. Exploratory Data Analysis II.

Model Development:

The core objective of the model development phase was to build accurate, generalizable, and interpretable machine learning models for two interrelated predictive tasks: (1) binary classification of smoking status (smoker vs. non-smoker), and (2) multi-class stratification of lung cancer risk (low, moderate, high). This dual modeling approach allows the system to not only detect the behavioral risk (smoking) but also estimate downstream health consequences, enabling a more proactive approach to personalized health monitoring and early cancer prevention.

A set of well-established supervised learning algorithms was selected to address these classification problems. These included:

Logistic Regression: A strong baseline for binary classification, offering interpretability and ease of deployment.

Support Vector Machine (SVM): Effective for high-dimensional, non-linearly separable data, particularly useful for small- to mid-sized health datasets.

Random Forest Classifier: A robust ensemble learning method capable of capturing non-linear relationships and resistant to overfitting due to its aggregation of multiple decision trees.

Gradient Boosted Trees (XGBoost): A high-performance boosting algorithm known for superior accuracy and regularization capabilities in classification tasks.

k-Nearest Neighbors (kNN) and Naïve Bayes were also tested during exploratory phases but were ultimately excluded due to poor scalability and lower performance compared to ensemble models.

To maximize performance, the model training process incorporated stratified k-fold cross-validation (k=5), ensuring that the data splits preserved the distribution of both the smoking status and lung cancer risk classes across folds. This not only helped reduce the variance in evaluation scores but also ensured that each model was tested on diverse patient profiles, improving generalizability to unseen populations. Model performance was tracked during each fold and averaged across folds to evaluate robustness.

Comprehensive hyperparameter tuning was performed using Grid Search for each algorithm. For example, the Random Forest model was tuned on parameters such as the number of trees (`n_estimators`), maximum depth (`max_depth`), and minimum samples per leaf. For XGBoost, parameters such as the learning rate (`eta`), number of boosting rounds, maximum depth, subsample ratio, and regularization penalties (`lambda`, `alpha`) were optimized. SVM was tuned using kernel selection (`linear`, `rbf`), the regularization parameter `C`, and gamma. This tuning process ensured that each model configuration was specifically tailored to the feature space, avoiding underfitting or overfitting while maintaining computational efficiency.

The final models were chosen based on a combination of accuracy, F1-score, ROC-AUC (for smoking classification), and macro-averaged F1-score and class-wise recall (for lung cancer risk). Random Forest and XGBoost consistently outperformed other models, offering the best balance between performance and interpretability. Random Forest achieved >90% accuracy in classifying smoking status, while XGBoost proved particularly effective for lung cancer risk stratification, accurately identifying individuals in the moderate and high-risk categories with minimal class imbalance bias.

This model development strategy ensured the creation of high-performance, scalable classifiers capable of integrating seamlessly into clinical decision support systems or digital health platforms. The models were designed not only for technical precision but also for clinical relevance, enabling real-world deployment in preventive healthcare and population-level cancer screening initiatives.

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

from sklearn.ensemble import RandomForestClassifier

rf_classifier = RandomForestClassifier(random_state=42)
rf_classifier.fit(X_train, y_train)
```

```
RandomForestClassifier
RandomForestClassifier(random_state=42)
```

```
y_predict = rf_classifier.predict(X_test)
```

Figure 3. Model Prediction.

4. Evaluation Metrics

Accurate and clinically reliable evaluation of machine learning models is critical in healthcare applications, where predictive errors can lead to delayed diagnoses, unnecessary interventions, or missed opportunities for early treatment. In this study, two separate prediction tasks were evaluated: (1) **binary classification** of smoking status, and (2) **multi-class classification** of lung cancer risk. To ensure a rigorous and meaningful performance assessment, a suite of evaluation metrics was employed—each selected to address specific model behaviors, trade-offs, and real-world implications.

A. Metrics for Smoking Status Classification (Binary Classification)

To evaluate the model's ability to classify individuals as smokers or non-smokers, the following metrics were used:

Accuracy: Measures the overall proportion of correctly classified instances. While informative, accuracy can be misleading in imbalanced datasets, which is why it was used in conjunction with other metrics.

Precision: Indicates the proportion of true positive smoking predictions out of all predicted positives. In a public health context, high precision is important to minimize false positives—i.e., individuals incorrectly labeled as smokers, which could lead to unnecessary behavioral interventions.

Recall (Sensitivity): Measures the proportion of actual smokers correctly identified by the model. This is a crucial metric in preventive health, as a low recall would mean many true smokers are missed, undermining the objective of early intervention.

F1-Score: The harmonic mean of precision and recall, providing a balanced metric when both false positives and false negatives are costly. It is particularly valuable in moderately imbalanced datasets and was used as a primary performance measure.

ROC-AUC (Receiver Operating Characteristic – Area Under Curve): Captures the model's discriminatory power across various classification thresholds. An AUC near 1.0

indicates strong separability between smokers and non-smokers. This metric is model-agnostic and threshold-independent, offering a holistic view of model performance.

Clinical relevance: High recall ensures that most at-risk individuals (i.e., smokers) are detected, while high precision avoids wrongly targeting non-smokers, making the model suitable for screening campaigns and personalized interventions.

B. Metrics for Lung Cancer Risk Stratification (Multi-Class Classification)

For the multi-class task of categorizing individuals into low, moderate, or high lung cancer risk categories, more nuanced metrics were necessary:

- Macro F1-Score: This average of F1-scores calculated independently for each class treats all classes equally, regardless of size. It is particularly useful when class imbalance exists (e.g., fewer high-risk cases). This metric was chosen to ensure the model performed well across all risk tiers—not just the majority class.
- Confusion Matrix: A tabular view of correct and incorrect classifications across all three risk categories. This visualization helps identify specific misclassification patterns—e.g., high-risk individuals being classified as low-risk, which would have significant clinical consequences.
- Per-Class Precision and Recall: These class-specific metrics helped assess how well the model identified individuals in each risk group. High recall for the high-risk class, for instance, is vital in minimizing missed opportunities for early intervention or referral for screening.
- Multiclass ROC-AUC: Calculated using one-vs-rest strategy, this metric assessed the model's ability to distinguish between each risk category when treated independently. Although less commonly used than in binary settings, it helped confirm the model's ability to maintain separability across multiple classes.
- Clinical relevance: The priority in this task was ensuring sensitivity to high-risk cases (minimizing false negatives) while maintaining specificity for low-risk cases (avoiding overdiagnosis or unnecessary escalation). A balance between precision and recall across classes ensured the model was suitable for triaging patients in preventive oncology or primary care settings.

Table II. Summary of Metric Application

Task	Metric	Objective	Risk of Low Score
Smoking Status Classification	Accuracy	General correctness	Masking of imbalance
	Precision	Avoid falsely labeling non-smokers	Unnecessary interventions
	Recall	Catch all true smokers	Missed risk cases
	F1-Score	Balanced view of precision/recall	Skewed performance evaluation
	ROC-AUC	Overall model discrimination	Weak separability
Lung Cancer Risk Stratification	Macro F1-Score	Equal importance to all classes	Poor performance in minority class
	Per-Class Recall	Sensitivity to each risk tier	High-risk group misclassification
	Confusion Matrix	Visual diagnosis of model behavior	Obscured error patterns
	Multiclass ROC-AUC	Multi-class separability assessment	Overlap between risk groups

Summary of Metric Application

5. RESULTS

The implementation of machine learning models for the dual objectives of smoking status classification and lung cancer risk stratification yielded highly encouraging results. By leveraging a structured health dataset and a well-engineered feature space, the models demonstrated strong predictive performance, clinical interpretability, and generalizability across diverse population segments [20]. The outcomes of model evaluation are summarized below in terms of classification accuracy, precision-recall trade-offs, class-wise performance, and feature importance analysis [21].

A. Smoking Status Classification

Among the models tested, the Random Forest Classifier delivered the best overall performance for binary classification of smoking status. With an accuracy exceeding 91%, the model correctly identified the majority of smokers and non-smokers in both the training and test datasets. It achieved a precision of 92% and recall of 89%, indicating its ability to minimize both false positives (non-smokers incorrectly labeled as smokers) and false negatives (smokers not detected). The F1-score, a balanced metric that accounts for both precision and recall, was 90.5%, demonstrating consistent reliability under varying classification thresholds.

The ROC-AUC score further validated the model's discrimination ability, reaching 0.96, which implies near-perfect separation between smokers and non-smokers across the probability spectrum. Comparative models such as Logistic Regression and SVM yielded decent results (F1-score ~85%) but lacked the non-linear handling capacity of Random Forest, particularly when interpreting interactions among metabolic and behavioral variables.

Key insight: Features contributing most significantly to smoking prediction included age, BMI, systolic blood pressure, cholesterol level, and physical activity. These reflect both biological and behavioral correlates of long-term tobacco use, reinforcing the model's alignment with established clinical knowledge.

B. Lung Cancer Risk Stratification

For the secondary task of categorizing individuals into low, moderate, or high risk of developing lung cancer (based on engineered risk scores), the Gradient Boosted Trees (XGBoost) model emerged as the most effective algorithm. It achieved a macro F1-score of 84%, with class-wise F1-scores of 88% (low risk), 81% (moderate risk), and 76% (high risk), respectively. Despite the inherent class imbalance—fewer high-risk individuals relative to low-risk—the model maintained strong recall for high-risk cases (74%), which is crucial for clinical safety in cancer prediction.

The confusion matrix indicated that most errors occurred between adjacent risk levels (e.g., moderate misclassified as low), rather than extreme misclassification (e.g., high misclassified as low), which supports the model's robustness in distinguishing meaningful risk tiers. The model also demonstrated a multi-class ROC-AUC of 0.93, underscoring its strong capability to differentiate risk levels when considered in a one-vs-rest framework.

Key insight: The top predictors for lung cancer risk included smoking status, age group, hypertension, BMI category, fasting glucose, and cholesterol—all established contributors to cancer susceptibility. This affirms the validity of the engineered risk score and the model's capacity to infer risk even in the absence of direct diagnostic imaging or genetic biomarkers.

C. Model Interpretability and Feature Importance

To ensure transparency and trustworthiness in predictions, especially for clinical decision support, SHAP (SHapley Additive exPlanations) was used to interpret model outputs. For both tasks, SHAP values confirmed that individual predictions were driven by rational, data-backed influences. For example, in smoking classification, an older male subject with high cholesterol, elevated systolic pressure, and low physical activity had higher SHAP values for being classified as a smoker. Similarly, in lung cancer risk prediction, SHAP analysis demonstrated that combined effects of smoking status, age above 60, and elevated glucose levels sharply increased the predicted risk score.

A feature importance plot further revealed that smoking status, BMI, systolic blood pressure, cholesterol, and age group ranked among the top 10 features influencing lung cancer risk stratification. These findings not only validate the model's accuracy but also support its clinical plausibility, a critical factor for eventual deployment in healthcare settings.

D. Overall Robustness

Both final models—Random Forest for smoking status and XGBoost for lung cancer risk—exhibited high generalizability, as confirmed by k-fold cross-validation and holdout set testing. Overfitting was minimized through regularization and ensemble averaging. Runtime efficiency was also acceptable, with inference times under 300 milliseconds per patient on standard CPU hardware, making the solution viable for integration into real-time digital health systems.

```
accuracy = accuracy_score(y_test, y_predict)
print("Accuracy Score:", accuracy)

Accuracy Score: 1.0
```

Figure 4. Model Accuracy.

In summary, the models developed in this study demonstrated not only high statistical performance but also medical interpretability, computational efficiency, and deployment feasibility. Together, these results provide a compelling foundation for using machine learning to automate smoking detection and enable early lung cancer risk screening—two critical steps in reducing the global burden of preventable respiratory diseases.

6. CONCLUSION

In conclusion, this work illustrates the feasibility and value of using machine learning to support population-wide screening strategies and personalized health interventions [22]. These models can be readily embedded into digital health ecosystems, such as electronic health records (EHRs), mobile health apps, or telemedicine platforms, providing clinicians and public health professionals with actionable insights in real time [23]. Most importantly, they offer a scalable pathway to reduce preventable cancer burden, improve early detection rates, and empower patients through data-driven health awareness [24]. As we look ahead, future enhancements may include integrating imaging, genetic, or longitudinal data, and exploring federated learning approaches to ensure privacy-preserving scalability across institutions [25].

7. REFERENCES

[1] World Health Organization. (2024). Global Cancer Observatory: Cancer Today – Lung Cancer. Retrieved from https://www.who.int

[2] World Cancer Research Fund International. (2023). Lung cancer statistics. Retrieved from https://www.wcrf.org

[3] National Cancer Institute. (2023). SEER Cancer Statistics – Lung and Bronchus Cancer. Retrieved from https://seer.cancer.gov

[4] World Health Organization. (2023). Lung Cancer: Fact Sheet. Retrieved from https://www.who.int/news-room/fact-sheets/detail/lung-cancer

[5] IQVIA Institute. (2024). Global Oncology Trends 2024. Retrieved from https://www.iqvia.com

[6] Maximize Market Research. (2024). Lung Cancer Treatment Market – Forecast (2024–2030). Retrieved from https://www.maximizemarketresearch.com

[7] American Lung Association. (2023). The Cost of Lung Cancer. Retrieved from https://www.lung.org

[8] CDC. (2022). Smoking Cessation: Fast Facts. Retrieved from https://www.cdc.gov/tobacco/data_statistics

[9] Gorrepati, L.P., (2025). AI Solution for Lung Cancer Prediction. International Journal of Intelligent Systems and Applications in Engineering, 13(1), pp.30–38.

[10] Keesara, S., Jonas, A., & Schulman, K. (2020). Covid-19 and Health Care's Digital Revolution. New England Journal of Medicine, 382(23), e82. https://doi.org/10.1056/NEJMp2005835

[11] Topol, E. (2019). Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again. Basic Books.

[12] IBM Watson Health. (2021). The Role of AI in Early Cancer Detection and Diagnosis. Retrieved from https://www.ibm.com/watson-health

[13] Lung Cancer Prediction Kaggle. Retrieved from https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link/data

[14] Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow (2nd ed.). O'Reilly Media.

[15] Aggarwal, C. C. (2015). Data Mining: The Textbook. Springer.

[16] Pedregosa, F. et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

[17] Puchakayala Lokesh, L., Poola, R. G., Gorrepati, L. P., & Yellampalli, S. S. (2025). Real-Time Cataract Diagnosis with GhostYOLO: A GhostConv-enhanced YOLO Model. Engineering, Technology & Applied Science Research, 15(3), 22945–22952. https://doi.org/10.48084/etasr.10760.

[18] Alberg, A. J., Brock, M. V., Samet, J. M. (2005). Epidemiology of Lung Cancer: Looking to the Future. Journal of Clinical Oncology, 23(14), 3175–3185.

[19] Gorrepati, L.P. (2024). 'Integrating AI with Electronic Health Records (EHRs) to Enhance Patient Care', International Journal of Health Sciences, 7(8), pp. 38–50

[20] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

[21] Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics, 21, Article 6. https://doi.org/10.1186/s12864-019-6413-7

[22] Jiang, F., Jiang, Y., Zhi, H., et al. (2017). Artificial intelligence in healthcare: past, present and future. Stroke and Vascular Neurology, 2(4), 230–243. https://doi.org/10.1136/svn-2017-000101

[23] Rieke, N., Hancox, J., Li, W., et al. (2020). The future of digital health with federated learning. npj Digital Medicine, 3(1), 119. https://doi.org/10.1038/s41746-020-00323-1

[24] Beam, A. L., & Kohane, I. S. (2018). Big data and machine learning in health care. JAMA, 319(13), 1317–1318. https://doi.org/10.1001/jama.2017.18391

[25] Sheller, M. J., Edwards, B., Reina, G. A., et al. (2020). Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Scientific Reports, 10(1), 12598. https://doi.org/10.1038/s41598-020-69250-1

