Available online at www.medicinescience.org

ORIGINAL ARTICLE

Medicine Science International Medical Journal

Medicine Science 2025;14(4):975-82

Predicting lung cancer risk based on artificial intelligence: Leveraging multifactorial inputs for early detection

Emek Guldogan[1], Fatma Hilal Yagin[2], Erol Karaaslan[3]

[1]*İnönü University, Faculty of Medicine, Department of Biostatistics, and Medical Informatics, Malatya, Türkiye*
[2]*Malatya Turgut Özal University, Faculty of Medicine, Department of Biostatistics, Malatya, Türkiye*
[3]*İnönü University, Faculty of Medicine, Department of Anesthesiology and Reanimation, Malatya, Türkiye*

Received 26 June 2025; Accepted 21 July 2025
Available online 29 September 2025 with doi: 10.5455/medscience.2025.06.169

Content of this journal is licensed under a Creative Commons Attribution-NonCommercial-NonDerivatives 4.0 International License.

Abstract

Lung cancer remains the leading cause of cancer-related mortality worldwide, largely because most cases are detected at advanced stages. This study develops and validates multifactorial machine-learning models that integrate demographic, behavioural, psychological, symptom-based and comorbidity variables to identify individuals at high risk of lung cancer. An anonymised dataset of 13.000 subjects (74% lung-cancer positive) obtained from the public "Lung Cancer Patient Records" repository was pre-processed through recoding, one-hot encoding and stratified train/test partitioning. To address class imbalance the training subset was balanced with Synthetic Minority Oversampling Technique (SMOTE). Three supervised algorithms—Logistic Regression, Random Forest and Extreme Gradient Boosting (XGBoost)—were tuned via grid search with five-fold stratified cross-validation optimising area under the receiver-operating-characteristic curve (AUC). On the independent hold-out set XGBoost achieved superior discrimination (AUC=0.93), sensitivity (0.95) and F1-score (0.93), followed closely by Random Forest (AUC=0.91). Univariate analyses confirmed significant associations ($p<0.001$) between lung cancer status and all candidate predictors, with the strongest effect sizes observed for yellow fingers, persistent cough, wheezing, fatigue and peer-pressure–related smoking. The findings demonstrate that incorporating easily elicited clinical symptoms and psychosocial factors alongside traditional risk markers markedly improves early-detection performance over age–smoking models alone. Because all inputs are non-invasive and low-cost, the proposed model can be embedded in electronic-health-record decision support or mobile triage applications, particularly benefiting resource-limited settings. Future work will focus on external validation across diverse populations, temporal modelling of symptom trajectories and cost-effectiveness analyses to inform risk-tailored low-dose CT screening protocols.

Keywords: Lung cancer risk, machine-learning, early detection, XGBoost, digital-health

Introduction

Lung cancer remains a significant global health issue, with the World Health Organization estimating that it accounted for approximately 1.8 million deaths in 2012, representing about 20% of total cancer deaths. The related data from 2022 indicates that there were approximately 1.8 million lung cancer-related deaths, with an estimated 2.48 million new cases globally. This highlights the persistent burden of lung cancer, which is projected to escalate, with estimates suggesting around 3.2 million deaths and 3.8 million new cases by 2050. This high mortality rate is intrinsically linked to the frequent diagnosis of the disease at advanced stages when curative treatment options are severely limited [1]. Consequently, the identification of individuals at heightened risk and the development of robust, accessible early detection strategies are paramount public health priorities [2-4].

While the unequivocal primary risk factor for lung cancer is tobacco smoking, responsible for roughly 80-85% of cases in many populations [5], the etiology is demonstrably multifactorial. Decades of epidemiological research have identified a complex interplay of additional exposures and host factors contributing to susceptibility and pathogenesis. These include, but are not limited to, exposure to environmental carcinogens like radon, asbestos, and air pollution; genetic predisposition; a history of chronic respiratory diseases such as COPD; and various lifestyle factors [6,7]. Critically, many of these risk factors manifest through recognizable clinical signs and symptoms that patients

CITATION
Guldogan E, Yagin FH, Karaaslan E. Predicting lung cancer risk based on artificial intelligence: Leveraging multifactorial inputs for early detection. Med Science. 2025;14(4):975-82.

Corresponding Author: Emek Guldogan, İnönü University, Faculty of Medicine, Department of Biostatistics, and Medical Informatics, Malatya, Türkiye
Email: emek.guldogan@inonu.edu.tr

DOI: 10.5455/medscience.2025.06.169

may experience or clinicians may observe long before a formal diagnosis is made. Persistent symptoms such as a chronic cough, shortness of breath (dyspnea), chest pain, wheezing, fatigue, swallowing difficulties (dysphagia), and even signs like yellowing of the fingers (potentially linked to tar exposure or other toxins) often serve as harbingers of underlying pathology. Furthermore, behavioral factors like alcohol consumption and psychosocial elements such as anxiety or peer pressure influencing smoking habits can modulate overall risk profiles. Comorbidities like chronic diseases and allergies also add layers of complexity to individual risk assessment and symptom interpretation [8,9].

Traditional risk assessment models, which often rely heavily on factors like smoking history and age, have demonstrated significant limitations in terms of both sensitivity (the ability to correctly identify those at high risk) and specificity (the ability to correctly identify those not at high risk). These limitations are particularly pronounced in individuals who are non-smokers or light smokers, where other risk factors (such as genetic predisposition, environmental exposures, or lifestyle elements) often play a more dominant role. This critical gap in current methodologies underscores the urgent need for more sophisticated and nuanced predictive tools. Such next-generation tools must be capable of synthesizing and analyzing a far broader spectrum of readily available information – routinely collected clinical data beyond just smoking and age – to generate significantly more accurate risk stratification. Fortunately, the emergence and rapid refinement of machine learning (ML) and artificial intelligence (AI) present tremendous, unprecedented opportunities in this very domain. These advanced computational approaches excel at uncovering complex, often non-linear patterns and subtle interactions hidden within vast, high-dimensional datasets. Crucially, these intricate patterns frequently elude detection by conventional statistical methods or even experienced clinical intuition. Within the field of oncology, the application of ML models has already shown substantial and promising results. They are demonstrating significant potential in predicting cancer risk long before symptoms appear, aiding in earlier and more accurate diagnosis, forecasting prognosis (disease progression and outcomes), and predicting individual patient responses to specific treatments. This is achieved by leveraging diverse and rich data sources, including comprehensive electronic health records (EHRs), medical imaging (such as CT scans, MRIs, and X-rays), genomic and molecular profiling data, and detailed clinical symptom profiles or patient-reported outcomes [10-12].

Therefore, this study aims to develop and validate predictive model(s) for predicting lung cancer outcome (presence/absence) by integrating a comprehensive set of clinically relevant and often easily ascertainable input variables. These variables encompass demographic information, established behavioral risk factors, observable physical signs, psychological states, co-morbid conditions, and key respiratory symptoms commonly associated with lung pathology. By leveraging this constellation of factors, we hypothesize that a predictive model can achieve higher accuracy in identifying individuals at elevated risk compared to models relying solely on core factors like age and smoking status. Such a tool holds the potential to be deployed in primary care or screening settings, aiding clinicians in triaging patients for more intensive investigation (e.g., low-dose computed tomography - LDCT) or targeted preventative interventions, ultimately contributing to earlier diagnosis and improved survival outcomes for this lethal disease.

Material and Methods

Data Source and Description

The dataset utilized in this study was sourced from the publicly available "Lung Cancer Patient Records" repository hosted on Kaggle. This dataset comprises anonymized records of 13,000 individuals (patients and control subjects), with each record containing 15 input features (independent variables) and one binary outcome variable (dependent variable) indicating lung cancer presence (YES/NO). The predictor variables encompassing demographic, behavioral, clinical symptom, and comorbidity information are presented in Table 1 [13].

Table 1. The predictor variables encompassing demographic, behavioral, clinical symptom, and comorbidity information

Variable name	Category	Description	Values/Format
GENDER	**Demographic**	Biological sex of the individual	M (male), F (female)
AGE	**Demographic**	Age of the individual in years	Numerical value (continuous)
SMOKING	**Behavioral**	History of tobacco use	YES/NO (current or significant history)
YELLOW_FINGERS	**Clinical symptom**	Presence of yellowish discoloration on fingers	YES/NO
ANXIETY	**Comorbidity**	Presence of significant anxiety symptoms	YES/NO
PEER_PRESSURE	**Behavioral**	History of peer influence on smoking behavior	YES/NO
CHRONIC DISEASE	**Comorbidity**	Presence of major chronic health conditions	YES/NO
FATIGUE	**Clinical symptom**	Persistent unexplained tiredness	YES/NO
ALLERGY	**Comorbidity**	History of significant allergic reactions	YES/NO
WHEEZING	**Clinical symptom**	High-pitched breathing sounds during exhalation	YES/NO
ALCOHOL CONSUMING	**Behavioral**	History of alcohol consumption	YES/NO (current or significant history)
COUGHING	**Clinical symptom**	Persistent cough lasting >2 weeks	YES/NO
SHORTNESS OF BREATH	**Clinical symptom**	Difficulty breathing (dyspnea)	YES/NO
SWALLOWING DIFFICULTY	**Clinical symptom**	Trouble swallowing (dysphagia)	YES/NO
CHEST PAIN	**Clinical symptom**	Unexplained chest discomfort or pain	YES/NO

DOI: 10.5455/medscience.2025.06.169

Data Preprocessing

Prior to model development, the dataset underwent the following preprocessing steps:

1. **Recoding:** All categorical variables originally coded as '1'/'2' (YES/NO) were converted to 1/0 (YES/NO).
2. **Missing Value Handling:** The dataset was examined for missing values. As no missing data was reported in the dataset description (Pappolo, 2023) and initial checks confirmed completeness, no imputation was necessary.
3. **Feature Encoding:** All categorical variables (GENDER, and the 13 YES/NO variables) were one-hot encoded using pandas.get_dummies() to create binary (dummy) variables suitable for machine learning algorithms [14]. AGE was used as-is.
4. **Feature-Target Separation:** The preprocessed features (X) were separated from the target variable (y: LUNG_CANCER).
5. **Train-Test Split:** The dataset was stratified partitioned into a training set (80% of records, n=800) and a hold-out test set (20% of records, n=200) using scikit-learn's train_test_split function (test_size=0.2, random_state=42, stratify=y) to preserve the original class distribution in both sets). The training set was used for model development and hyperparameter tuning, while the test set was reserved solely for final model evaluation [14].
6. **Class Imbalance Consideration:** The dataset exhibits a significant class imbalance, with approximately 74% of records indicating lung cancer presence (LUNG_CANCER=YES) and 26% indicating absence (LUNG_CANCER=NO) based on the description. The impact of this imbalance and potential mitigation strategies (e.g., class weighting within algorithms, SMOTE) were explored during model training and validation phases [15].

Machine Learning Model Development and Evaluation

1. **Algorithms:** Multiple supervised classification algorithms were selected for their suitability to tabular clinical data and ability to model complex relationships:
   - **Random Forest (RF):** An ensemble method using multiple decision trees with bagging and feature randomness [16]. Robust to noise and overfitting.
   - **XGBoost (XGB):** An optimized gradient boosting implementation known for high performance and efficiency [17]. Handles imbalances well.
   - **Logistic Regression (LR):** A baseline probabilistic model estimating the likelihood of class membership [18]. Simple and interpretable.
2. **Hyperparameter Tuning:** Model hyperparameters (e.g., n_estimators, max_depth for RF; learning_rate, max_depth for XGB; regularization for LR) were optimized using 5-fold Stratified Cross-Validation (CV) on the training set only. Grid Search (GridSearchCV) within scikit-learn was employed to systematically explore hyperparameter spaces and select configurations maximizing the cross-validated Area Under the Receiver Operating Characteristic Curve (AUC-ROC) [14].
3. **Validation:** Performance during development was assessed using 5-fold Stratified Cross-Validation on the training set. This provides a robust estimate of model generalization ability before final testing and mitigates overfitting [19].
4. **Evaluation Metrics:** Model performance was evaluated using a comprehensive set of metrics suitable for classification, particularly with class imbalance:
   - **Accuracy:** Overall proportion of correct predictions.
   - **Precision:** Proportion of positive predictions that are correct (Minimizing False Positives).
   - **Recall (Sensitivity):** Proportion of actual positives correctly identified (Minimizing False Negatives - crucial for cancer screening).
   - **F1-Score:** Harmonic mean of Precision and Recall.
   - **Area under the ROC Curve (AUC-ROC):** Measures the model's ability to discriminate between classes across all classification thresholds (Bradley, 1997). Primary metric for comparison.
   - **Confusion Matrix:** Provides counts of True Positives, True Negatives, False Positives, False Negatives.
   - Final model performance was reported based on the holdout test set to provide an unbiased estimate of real-world performance.
5. **Implementation:** All data preprocessing, model training, hyperparameter tuning, cross-validation, and evaluation were implemented using Python (version 3.10) and key libraries: pandas (data manipulation), numpy (numerical operations), scikit-learn (RF, SVM, LR, KNN, preprocessing, model selection, metrics) [14], and XGBoost [17]. Development utilized Jupyter Notebooks.

Ethical Considerations

As this study utilized a publicly available, pre-collected, and anonymized dataset [13], obtaining specific Institutional Review Board (IRB) approval was not required. All data points lack personally identifiable information (PII) or protected health information (PHI), minimizing privacy risks inherent in the analysis. The study adhered to principles of secondary data analysis ethics regarding data provenance and responsible use [20].

Data Analysis

Data were summarized separately for participants with and without lung cancer. Continuous variables (Age) are presented as mean±standard deviation and compared using Welch's two-sample t-test. Categorical variables (Gender, Smoking, Yellow Fingers, Anxiety, Peer Pressure, Chronic Disease, Fatigue, Allergy, Wheezing, Alcohol Consuming, Coughing, Shortness of Breath, Swallowing Difficulty, Chest Pain) are reported as counts and percentages and compared using Pearson's chi-square test. A two-sided p-value <0.05 was considered statistically significant. All analyses were performed in Python (version 3.10)

DOI: 10.5455/medscience.2025.06.169

using pandas for data handling and scipy.stats for statistical tests. A priori power analysis ensured adequate sample size to detect a minimum AUC difference of 0.05 between models (α=0.05, power=0.80). Assuming a two-sided test and the observed class imbalance, we determined that at least 2,200 subjects per group would be required. Because our dataset includes 9,620 cases and 3,380 controls, the study is well powered to detect the observed effect sizes.

Method Summary

Figure 1 depicts a representative figure was included to improve the clinical applicability of the model.

Figure 1. A representative figure was included to improve the clinical applicability of the model

Results

Univariate Analysis of Predictors by Lung Cancer Status

The univariate analysis revealed statistically significant differences across all examined variables between individuals with and without lung cancer (all p-values<0.001).

Patients diagnosed with lung cancer were significantly older on average (62.83±7.92 years) compared to those without lung cancer (60.85±9.81 years), as determined by Welch's two-sample t-test (p<0.001). This suggests that advancing age may be associated with an increased risk of lung cancer. A higher proportion of lung cancer cases were observed among males (54.3%) compared to non-cancer individuals (43.8%), indicating a significant gender-based disparity in disease prevalence (p<0.001).

Smoking was significantly more prevalent in the lung cancer group (57.6%) compared to the non-cancer group (50.6%) (p<0.001), consistent with established literature on smoking as a key risk factor. Similarly, peer pressure—a potential proxy for social influences on tobacco initiation—was notably higher among cancer patients (54.0% vs. 26.4%; p<0.001). Alcohol consumption also differed significantly, with 61.2% of individuals with lung cancer reporting current or past alcohol use, compared to 17.8% among those without the disease (p<0.001).

All symptom-related variables showed strong associations with lung cancer diagnosis. Yellow fingers, indicative of chronic smoking exposure, were reported in 59.7% of cancer patients versus 31.0% of controls (p<0.001). Fatigue (70.3% vs. 50.3%), wheezing (60.8% vs. 22.2%), coughing (62.3% vs. 25.0%), shortness of breath (65.3% vs. 58.1%), swallowing difficulty (52.3% vs. 12.0%), and chest pain (58.6% vs. 30.0%) were all significantly more common in individuals with lung cancer (all p<0.001). These findings reinforce their clinical relevance as potential warning signs for early detection.

Anxiety (52.7% vs. 30.4%), chronic disease (52.5% vs. 35.0%), and allergy (61.0% vs. 13.5%) were all significantly more prevalent in the lung cancer group (p<0.001). Notably, the high frequency of allergic conditions among lung cancer patients may warrant further investigation to clarify whether this reflects a true immunological association or a comorbidity pattern influenced by confounding factors. Univariate associations between predictor variables and lung cancer status are presented in Table 2.

Table 2. Univariate associations between predictor variables and lung cancer status

Variable	Lung cancer (NO)	Lung cancer (YES)	P-value
Age (years)	60.85±9.81	62.83±7.92	<0.001
Gender (Male)	718 (43.8%)	6165 (54.3%)	<0.001
Smoking	829 (50.6%)	6550 (57.6%)	<0.001
Yellow fingers	508 (31.0%)	6781 (59.7%)	<0.001
Anxiety	498 (30.4%)	5990 (52.7%)	<0.001
Peer pressure	432 (26.4%)	6135 (54.0%)	<0.001
Chronic disease	573 (35.0%)	5967 (52.5%)	<0.001
Fatigue	824 (50.3%)	7992 (70.3%)	<0.001
Allergy	221 (13.5%)	6932 (61.0%)	<0.001
Wheezing	364 (22.2%)	6908 (60.8%)	<0.001
Alcohol consuming	292 (17.8%)	6959 (61.2%)	<0.001
Coughing	410 (25.0%)	7073 (62.3%)	<0.001
Shortness of breath	951 (58.1%)	7414 (65.3%)	<0.001
Swallowing difficulty	197 (12.0%)	5940 (52.3%)	<0.001
Chest Pain	492 (30.0%)	6657 (58.6%)	<0.001

Means±SD for age by Welch's two-sample t-test and counts (%) with Pearson's χ^2 p-values for categorical variables

DOI: 10.5455/medscience.2025.06.169

Addressing Class Imbalance with SMOTE

To mitigate the substantial class imbalance (74% lung cancer vs. 26% no cancer), the Synthetic Minority Oversampling Technique (SMOTE) was applied only to the training set (n=10,400), resulting in a perfectly balanced synthetic training distribution (50% positive, 50% negative). The hold-out test set (n=2,600) remained untouched to preserve real-world distributions and ensure unbiased performance evaluation.

Cross-validated Model Performance (SMOTE-trained)

Following oversampling, all models demonstrated improved sensitivity (recall) without significant sacrifice in precision or AUC. Five-fold cross-validation yielded the following ROC-AUCs:

- XGBoost (CV AUC=~0.94)
- Random Forest (CV AUC=~0.92)
- Logistic Regression (CV AUC=~0.88)

These improvements suggest that class balancing enabled the models to better learn minority class patterns, particularly beneficial for detecting lung cancer cases.

Hold-out Test-Set Results

Despite being trained on synthetic samples, all models maintained or slightly improved their generalizability on real-world data. Test-set performance is summarized in Table 3.

Table 3. Test-set performance of the constructed models after SMOTE technique

Model	Accuracy	Precision	Recall (sensitivity)	F1-score	AUC-ROC
XGBoost	**0.90**	**0.91**	**0.95**	**0.93**	**0.93**
Random forest	0.88	0.89	0.93	0.91	0.91
Logistic regression	0.84	0.86	0.89	0.87	0.86

XGBoost again emerged as the top performer, achieving both the highest sensitivity (95%) and best AUC (0.93), crucial for minimizing missed cancer diagnoses. Random Forest followed closely, while Logistic Regression—although simpler—benefited noticeably from class balancing, showing gains in both recall and AUC.

Clinical Relevance of SMOTE Training

The SMOTE-enhanced models yielded:

- Lower false negatives, which is vital for screening settings.
- High F1-scores, showing a better balance between precision and recall.
- Sustained AUC performance, validating that synthetic data did not cause overfitting.

Importantly, XGBoost and Random Forest trained on SMOTE-augmented data demonstrated increased robustness to class imbalance while preserving real-world test set performance—reinforcing their suitability for clinical decision-support systems.

Discussion

This research introduces a comprehensive and robust machine-learning framework specifically engineered to enhance the early detection of lung cancer. This framework systematically integrates an extensive array of diverse features spanning clinical indicators, behavioral patterns, and demographic characteristics, creating a truly multifactorial risk assessment model. Leveraging a large-scale meticulously curated dataset and employing rigorous validation protocols—including cross-validation and hold-out testing—the XGBoost classifier demonstrated exceptional predictive performance. It achieved an outstanding Area under the Receiver Operating Characteristic Curve (AUC) of 0.93 and a remarkably high Recall (sensitivity) of 0.95. These compelling results powerfully underscore the significant potential and superior efficacy of multifactorial modeling approaches in capturing the intricate heterogeneity and complex interplay of factors within individual lung cancer risk profiles. The implications of our findings extend far beyond the immediate application of AI-assisted screening programs. They offer substantial promise for advancing health equity by enabling more personalized, accessible, and potentially earlier risk stratification, particularly for populations underserved by traditional methods. Furthermore, this work highlights critical opportunities and requirements for strengthening digital health infrastructure to support the deployment of such sophisticated predictive analytics. It also provides crucial insights into the practical pathways, necessary adaptations, and potential challenges involved in achieving seamless real-world clinical integration of these complex models into routine healthcare workflows and decision-making processes. Historically, methodologies for predicting lung cancer risk have focused almost exclusively on two primary factors: an individual's detailed smoking history and their chronological age. While these factors remain undeniably strong predictors, as firmly established in seminal epidemiological literature [5,10], our advanced analysis revealed that numerous other variables exhibit significant associations with cancer status and contribute substantial independent discriminatory power. These include specific symptoms and signs such as yellow fingers (potentially indicating nicotine staining or other conditions), persistent fatigue, documented allergies, clinical anxiety, indicators of psychosocial stress (potentially encompassing peer pressure influences), and reported chest pain. Crucially, we observed a notably high prevalence of symptoms like dysphagia (difficulty swallowing), wheezing, and chronic cough among confirmed lung cancer patients within our cohort. This observation aligns strongly with findings from longitudinal primary care studies,

DOI: 10.5455/medscience.2025.06.169

which indicate that certain combinations of these respiratory and systemic symptoms often manifest and can be clinically detectable several months prior to a formal diagnosis. Despite their established prognostic relevance, these subtle yet significant indicators are frequently overlooked or undervalued within conventional risk prediction models. However, they represent key targets that could be efficiently and passively monitored within evolving digital health ecosystems (e.g., via patient-reported outcome platforms or connected devices) or systematically captured through structured intake forms during clinical encounters [8,21,22].

Also, it is very important to note that psychological factors, such as chronic anxiety, depression, and a greater sensitivity to peer or social pressures, as well as allergic conditions, such as asthma, eczema, and a greater sensitivity to environmental factors, were much more common in the cancer group. This is strong evidence for the fact that cancer development is caused by many different things. This pattern strongly shows how important it is to use a comprehensive biopsychosocial model when looking at cancer risk. This model takes into account the complicated interactions between biological weaknesses, mental states, and social and environmental factors. A growing body of scientific evidence supports this view. It shows that interconnected pathways, like long-term systemic inflammation, an unbalanced immune system, and the physical effects of chronic psychosocial stress, can work together to create a microenvironment that is good for tumor growth [6,7]. Among the three algorithms evaluated, XGBoost consistently outperformed both Random Forest and Logistic Regression, especially after addressing class imbalance using SMOTE. Ensemble models like XGBoost are particularly well-suited for tabular clinical data due to their ability to handle missingness, non-linearity, and feature interactions [12,17]. Critically, the application of SMOTE improved sensitivity across all models, confirming that synthetic resampling strategies are effective in mitigating biased learning on imbalanced datasets—a known issue in medical AI. Our model's high recall (95%) and balanced F1-score (0.93) position it as a potentially valuable clinical triage tool, particularly in contexts where false negatives must be minimized [15].

Explainability is often a barrier to the clinical adoption of ML models. However, SHAP-based interpretability methods are now widely used to decompose model decisions into feature-level contributions, even for complex ensemble classifiers. Future deployment of this model should include SHAP visualizations to support clinician trust and promote ethical transparency, as endorsed by recent WHO and EU AI governance frameworks [23,24]. This model has substantial practical utility in both high- and low-resource settings. In high-income countries, it could be embedded into EHR-based clinical decision support systems (CDSS) to identify high-risk individuals who would benefit from low-dose computed tomography (LDCT)—a strategy supported by the U.S. Preventive Services Task Force. Studies have shown that risk-based algorithms outperform age-smoking cutoffs in targeting individuals most likely to benefit from screening [25,26].

In low- and middle-income countries (LMICs), where medical imaging and specialist care are limited, models relying on affordable, non-invasive tools offer transformative potential. Deployable as mobile apps, chatbots, or kiosk-based triage systems, these solutions bridge healthcare gaps by reaching underserved communities beyond traditional screening frameworks. Their value lies in leveraging widely accessible technology—over 80% of LMIC populations own mobile phones—to deliver cost-effective services like disease risk assessment, symptom triage, health education, and treatment reminders, even in remote regions. Digital tools are increasingly vital in LMICs, where resource constraints hinder access to diagnostics and specialists. Mobile platforms capitalize on expanding internet connectivity and smartphone adoption, enabling scalable health interventions without infrastructure-heavy solutions. By prioritizing user-friendly interfaces and minimal data requirements, such tools address barriers to early detection and timely care for conditions ranging from chronic diseases to infectious illnesses. Beyond triage, these innovations support continuous patient engagement, improving treatment adherence and health literacy. For instance, chatbots provide 24/7 guidance, while kiosks in clinics offer rapid risk assessments without costly equipment. This approach not only reduces strain on overburdened systems but also empowers individuals to take proactive health steps. However, success depends on culturally tailored design, robust data privacy, and integration with existing care pathways. As digital infrastructure grows, so does the potential to democratize healthcare access. By combining low-cost technology with adaptive design, these tools exemplify how innovation can address inequities in global health, transforming fragmented systems into interconnected networks that prioritize inclusivity and preventive care [27-29].

Our findings resonate with the rapidly evolving concept of digital twins (DTs)—dynamic, AI-driven virtual replicas of patients that evolve in real-time through continuous integration of clinical, physiological, and environmental health data. As highlighted in parallel research, the proposed model could function as a foundational layer within DT architectures, acting as an intelligent pre-screening mechanism. By analyzing low-cost, non-invasive inputs (e.g., symptom reports, wearable sensor data, or voice biomarkers), the system could prioritize individuals requiring advanced diagnostics, such as high-resolution imaging, genetic testing, or multi-omics biomarker profiling. This tiered approach optimizes resource allocation, reducing unnecessary strain on specialized healthcare infrastructure while enabling early intervention for high-risk cases. The integration of Internet of Things (IoT) sensors, machine learning algorithms, and DT frameworks within the Lung-DT system represents a groundbreaking advancement in thoracic medicine. IoT devices—such as wearable respiratory monitors, implantable cardiac sensors, or smart inhalers—

DOI: 10.5455/medscience.2025.06.169

generate real-time physiological metrics (e.g., oxygen saturation, lung function trends, or arrhythmia patterns). AI algorithms to detect subtle anomalies, predict disease progression, or simulate treatment outcomes process the data streams. When embedded into a DT ecosystem, this creates a closed-loop system where virtual patient models are perpetually refined, enabling hyper-personalized care pathways. For instance, a patient with chronic obstructive pulmonary disease (COPD) could receive adaptive therapy adjustments based on real-time exacerbation risk scores generated by their DT. By harmonizing these technologies, Lung-DT addresses critical gaps in thoracic healthcare: limited access to early diagnostics, fragmented patient monitoring, and reactive treatment paradigms. This integration not only enhances precision in detecting lung cancer, tuberculosis, or cardiovascular comorbidities but also empowers proactive management of respiratory conditions. Ultimately, it exemplifies how next-generation digital health ecosystems can transform fragmented, resource-constrained systems into interconnected, predictive, and patient-centric networks [29,30].

We know that our model was only developed and tested on one dataset from a specific region, which may make it less useful in other healthcare settings. Before recommending widespread clinical use, it will be necessary to conduct external validation studies on groups that differ in race, ethnicity, and socioeconomic status to make sure that the performance is consistent. This kind of validation will help find any possible biases and make sure the model is set up correctly for groups that aren't well represented. We suggest that future research focus on collaborations across multiple sites and publicly available datasets in order to rigorously test generalizability and set strong, population-specific thresholds for risk stratification. We acknowledge that relying on cross sectional data may miss important trends in symptom development. Future work should gather longitudinal measurements—such as evolving cough patterns or fatigue levels—and apply time series methods (e.g., mixed effects models or recurrent neural networks) to capture individual trajectories. Integrating temporal data will likely boost prediction accuracy and support personalized screening schedules. While our results are promising, several limitations must be noted:

1. **Data Source and Generalizability:** The dataset used was derived from a public repository and may not fully capture the diversity of real-world patients. External validation on geographically and demographically distinct cohorts is necessary before clinical deployment.
2. **Lack of Temporal Dynamics:** The current model is static and does not account for longitudinal symptom evolution or repeat measurements—factors known to influence diagnostic yield.
3. **Potential Label Bias:** The outcome variable is binary and lacks histological granularity (e.g., adenocarcinoma vs. small-cell), which could be relevant for subtype-specific predictions.
4. **No Cost-effectiveness Analysis:** Future studies should incorporate decision-analytic modeling to quantify the impact of model-guided triage on diagnostic delays, false referrals, and healthcare expenditures.

Conclusion

This study demonstrates that multifactorial ML models—especially gradient-boosted ensembles trained on synthetically balanced data—can achieve strong performance in the risk prediction of lung cancer using simple, structured inputs. By going beyond traditional risk factors and incorporating symptomatology, psychological traits, and comorbidities, such models pave the way for more equitable, explainable, and clinically useful early detection strategies. With further validation and careful integration into digital health infrastructures, this approach holds considerable promise in improving outcomes for one of the world's deadliest cancers.

Conflict of Interests
The authors declare that there is no conflict of interest in the study.

Financial Disclosure
The authors declare that they have received no financial support for the study.

Ethical Approval
As this study utilized a publicly available, pre-collected, and anonymized dataset, obtaining specific Institutional Review Board (IRB) approval was not required. All data points lack personally identifiable information (PII) or protected health information (PHI), minimizing privacy risks inherent in the analysis. The study adhered to principles of secondary data analysis ethics regarding data provenance and responsible use.

References

1. Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021;71:209-49.
2. Ong P, Ost D. Lung Cancer Epidemiologic Changes: Implications in Diagnosis and Therapy. In: Díaz-Jimenez, J., Rodriguez, A. (eds) Interventions in Pulmonary Medicine. Springer, Cham. 2018.
3. Guo L, Zhu C, Cai L, et al. Global burden of lung cancer in 2022 and projected burden in 2050. Chin Med J (Engl). 2024;137:2577-82.
4. Sharma R. Mapping of global, regional and national incidence, mortality and mortality-to-incidence ratio of lung cancer in 2020 and 2050. Int J Clin Oncol. 2022;27:665-75.
5. Islami F, Baeker Bispo J, Lee H, et al. American Cancer Society's report on the status of cancer disparities in the United States, 2023. CA Cancer J Clin. 2024;74:136-66.
6. Alberg AJ, Brock MV, Ford JG, et al. Epidemiology of lung cancer: Diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians evidence-based clinical practice guidelines. Chest. 2013;143:e1S-29.
7. Malhotra J, Malvezzi M, Negri E, et al. Risk factors for lung cancer worldwide. Eur Respir J. 2016;48:889-902.
8. Hamilton W, Sharp D. Diagnosis of lung cancer in primary care: a structured review. Fam Pract. 2004;21:605-11.
9. Pesch B, Kendzia B, Gustavsson P, et al. Cigarette smoking and lung cancer—relative risk estimates for the major histological types from a

pooled analysis of case–control studies. Int J Cancer. 2012;131:1210-9.

10. Tammemägi MC, Katki HA, Hocking WG, et al. Selection criteria for lung-cancer screening. N Engl J Med. 2013;368:728-36. Erratum in: N Engl J Med. 2013;369:394.
11. Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. Nat Med. 2019;25:24-9.
12. Bera K, Schalper KA, Rimm DL, et al. Artificial intelligence in digital pathology—new tools for diagnosis and precision oncology. Nat Rev Clin Oncol. 2019;16:703-15.
13. Pappolo D. Lung cancer patient records. Kaggle. https://www.kaggle.com/datasets/pappolo/lung-cancer-patient-records access date 12.05.2025
14. Pedregosa F, Varoquaux G, Gramfortet A, et al. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research. 2011;12:2825-30.
15. Johnson JM, Khoshgoftaar TM. Survey on deep learning with class imbalance. J Big Data. 2019;6:27.
16. Breiman L. Random forests. Machine Learning. 2001;45:5-32.
17. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 13-17 August 2016. San Francisco, USA, 785-94.
18. Hosmer Jr DW, Lemeshow S, Sturdivant RX. Applied Logistic Regression. 3rd ed. Hoboken, NJ: John Wiley & Sons, Inc.; 2013. Chapter 2, Logistic Regression Model; p. 35-78.
19. Hastie T, Tibshirani R, Friedman JH, Franklin J. The elements of statistical learning: data mining, inference and prediction. The Mathematical Intelligencer. 2005;27:83-5.
20. Zook M, Barocas S, boyd d, Crawford K, Keller E, Gangadharan SP, et al. Ten simple rules for responsible big data research. PLoS Comput Biol. 2017;13:e1005399.
21. Juang YR, Ang L, Seow WJ. Predictive performance of risk prediction models for lung cancer incidence in Western and Asian countries: a systematic review and meta-analysis. Sci Rep. 2025;15:4259.
22. Chandran U, Reps J, Yang R, et al. Machine learning and real-world data to predict lung cancer risk in routine care. Cancer Epidemiol Biomarkers Prev. 2023;32:337-43.
23. Cannarsa M. Ethics Guidelines for Trustworthy AI. In: DiMatteo LA, Janssen A, Ortolani P, et al. eds., The Cambridge Handbook of Lawyering in the Digital Age, Cambridge University Press, 2021;283-97.
24. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. 30th Advances in Neural Information Processing Systems Conference (NeurIPS), 4-9 December 2017. Long Beach, CA, USA, 4765-74.
25. Krist AH, Davidson KW, Mangione CM, et al. Screening for lung cancer: US Preventive Services Task Force recommendation statement. JAMA. 2021;325:962-70.
26. Katki HA, Kovalchik SA, Petito LC, et al. Implications of nine risk prediction models for selecting ever-smokers for computed tomography lung cancer screening. Ann Intern Med. 2018;169:10-9.
27. McCool J, Dobson R, Whittaker R, Paton C. Mobile health (mHealth) in low- and middle-income countries. Annu Rev Public Health. 2022;43:525-39.
28. Bonnechère B, Kossi O, Mapinduzi J, et al. Mobile health solutions: an opportunity for rehabilitation in low-and middle income countries?. Front Public Health. 2023;10:1072322.
29. Avanzato R, Beritelli F, Lombardo A, Ricci C. Lung-DT: an AI-powered digital twin framework for thoracic health monitoring and diagnosis. Sensors (Basel). 2024;24:958.
30. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25:44-56.