Cancer Screening and Prevention **2026** | Epub
DOI: 10.14218/CSP.2026.00009

Original Article

Explainable Ensemble Machine Learning Framework for Lung Cancer Status Classification: A Retrospective Cross-sectional Study

Hamza Saad*

School of Industrial Sciences and Technology, University of Central Missouri, Warrensburg, MO, USA

Received: June 06, 2026 | **Revised:** June 29, 2026 | **Accepted:** August 14, 2026 | **Published online:** September 08, 2026

Abstract

Background and objective: Machine-learning approaches that combine predictive performance with model interpretability may improve lung cancer status classification. This study aimed to develop and internally evaluate an Explainable Precision Screening Framework for lung cancer risk classification using demographic, behavioral, and symptom-based variables from a publicly available dataset.

Methods: This retrospective cross-sectional study analyzed a publicly available Kaggle dataset containing 309 records, including 270 labeled as lung cancer and 39 as non-cancer. Six models—logistic regression, support vector machine (SVM), random forest, LightGBM, XGBoost, and a stacking ensemble—were compared using a stratified hold-out test set and repeated stratified five-fold cross-validation. Performance was assessed using classification metrics with bootstrap 95% confidence intervals (CIs). A separate Shapley additive explanations (SHAP) analysis was applied to the standalone SVM model for exploratory feature attribution.

Results: Across all models, accuracy ranged from 85% to 92% (ROC-AUC: 0.93–0.95). The stacking ensemble achieved 0.92 accuracy (95% CI: 0.85–0.98), 0.94 sensitivity (95% CI: 0.88–1.00), 0.75 specificity (95% CI: 0.40–1.00), 0.96 precision (95% CI: 0.89–1.00), 0.95 F1 score (95% CI: 0.91–0.99), and 0.95 ROC-AUC (95% CI: 0.89–0.99). Its PR-AUC was 0.993 (95% CI: 0.981–0.999), and its Brier score was 0.074 (95% CI: 0.037–0.121). SHAP analysis identified smoking, yellow fingers, coughing, chest pain, wheezing, shortness of breath, and age as the features contributing most strongly to its predictions.

Conclusions: Within this public retrospective dataset, the stacking ensemble was among the highest-performing models, whereas separate SHAP analysis of the standalone SVM model identified the features contributing most strongly to its predictions. Given the marked class imbalance, the absence of a reported clinical reference standard for the source labels, the framework requires independent validation in clinically verified multicenter cohorts before clinical use.

Introduction

Lung cancer remains the leading cause of cancer-related death worldwide, accounting for approximately 1.8 million deaths annually and imposing a substantial global public health burden despite advances in prevention, diagnosis, and treatment.[1,2] Many patients are diagnosed at an advanced stage, limiting curative treatment options. Data indicate that 5-year survival exceeds 60% for disease diagnosed at an earlier stage but is < 10% for metastatic disease.[1] Accordingly, developing reliable, accessible, and cost-effective approaches to earlier detection remains an important goal in precision oncology.

Low-dose computed tomography (LDCT) screening reduces lung cancer mortality in populations at elevated risk. However, implementation may be limited by cost, access, radiation exposure, false-positive results, and suboptimal adherence.[3-5] Existing risk-assessment tools often rely on a limited set of smoking-related and demographic variables and may not fully

Keywords: Lung cancer; Machine learning; Ensemble learning; Explainable artificial intelligence (XAI); SHAP; Classification.
***Correspondence to:** Hamza Saad, School of Industrial Sciences and Technology, University of Central Missouri, Warrensburg, MO, USA. ORCID: https://orcid.org/0000-0002-9647-6833. Tel: +1-660-543-4733, E-mail: saad@ucmo.edu
How to cite this article: Saad H. Explainable Ensemble Machine Learning Framework for Lung Cancer Status Classification: A Retrospective Cross-sectional Study. *Cancer Screen Prev*. Published online: Sep 07, 2026. DOI: https://doi.org/10.14218/CSP.2026.00009.

capture interactions among behavioral, symptom-related, and clinical factors. Artificial intelligence (AI) and machine learning (ML) may help integrate multidimensional data to support individualized risk estimation and lung cancer screening.[6]

Numerous studies have shown that machine-learning algorithms, including random forest, support vector machine (SVM), XGBoost, and deep-learning models, can perform well in cancer prognosis and prediction.[7,8] Recent work emphasizes that clinical adoption of AI requires not only predictive performance but also transparency, explainability, rigorous validation, and integration into clinical workflows.[9] However, many existing models prioritize predictive accuracy over interpretability, which can limit clinician trust and clinical implementation. In addition, relatively few studies have applied explainable artificial intelligence (XAI) methods to characterize features contributing to lung cancer status classification.[10] These limitations hinder evaluation of transparent, explainable AI frameworks in oncology.

This study develops the Explainable Precision Screening Framework (EPSF), which evaluates ensemble machine-learning algorithms for lung cancer status classification and uses Shapley additive explanations (SHAP) separately for exploratory interpretation of a standalone SVM model. The EPSF uses demographic, behavioral, and symptom variables to generate predictions and characterize model-based feature contributions. The framework is evaluated as an exploratory research model and requires external clinical validation.

This study aimed to develop and evaluate an ensemble machine-learning model for lung cancer status classification. The primary research question was whether a stacking ensemble incorporating XGBoost, LightGBM, and random forest would outperform individual machine-learning classifiers; a complementary SHAP analysis was performed separately on the standalone SVM model.[11-13]

To address this objective, a retrospective dataset containing records labeled as lung cancer or non-cancer was analyzed. The data were preprocessed, correlation analyses were performed, predictive models were developed using established machine-learning algorithms, cross-validation was conducted, and model performance was evaluated. Performance metrics included accuracy, precision, recall (sensitivity), F1 score, area under the receiver operating characteristic curve (ROC-AUC), area under the precision-recall curve (PR-AUC), and Brier score. A separate SHAP analysis of the standalone SVM model was used to characterize behavioral and symptom-related feature contributions.[14]

This study was designed to contribute in three ways: introducing an explainable machine-learning framework for lung cancer status classification; identifying the variables that contributed most strongly to predictions within this dataset; and providing an exploratory basis for future validation of transparent AI models in clinically verified cohorts.

AI is increasingly applied to lung cancer research because of the disease's high mortality and the potential benefits of earlier detection. Many studies have used ML to identify high-risk groups from demographic, behavioral, imaging, genetic, and clinical predictors.[10,15] These efforts align with precision oncology, which tailors screening and prevention strategies to individual risk profiles rather than relying solely on population-level criteria.

Logistic regression is widely used for cancer risk estimation because of its relatively simple structure and interpretability.[16,17] Studies have used logistic regression to identify important risk factors, including smoking status, age, and chronic lung disease. Its principal limitation is reduced flexibility for modeling nonlinear relationships and complex interactions among predictors.[7] Consequently, machine-learning methods have increasingly been used to model complex patterns in healthcare data.

Random forest is a widely used ensemble method for cancer prediction. Breiman's approach improves predictive stability and can reduce overfitting by aggregating predictions across multiple decision trees.[11] Studies have reported good performance of random-forest models for lung cancer risk classification using clinical and behavioral information.[10] However, random forest is generally less directly interpretable than traditional statistical models, which can complicate clinical implementation.

Gradient-boosting algorithms such as XGBoost and LightGBM have attracted substantial attention because of their performance on structured medical datasets.[13] XGBoost can efficiently model nonlinear relationships and interactions while maintaining computational scalability. LightGBM similarly offers efficient training while retaining strong predictive performance.[12] Studies in oncology have reported that boosting algorithms can outperform traditional machine-learning approaches for some cancer-prediction tasks.[7] Nevertheless, many studies emphasize predictive accuracy while providing limited insight into the clinical relevance of identified predictors.

Neural networks can achieve strong performance in radiologic image analysis and genomic data analysis.[6] However, deep-learning models can be prone to overfitting in small clinical datasets and may require substantial computational resources; their limited interpretability can also hinder use in routine clinical decision support, particularly where transparency is important for regulatory and ethical reasons.

XAI has emerged to address the limited transparency of "black-box" machine-learning models. Lundberg and Lee developed SHAP to quantify the contribution of individual features to model predictions.[18] SHAP is increasingly used in healthcare research to improve transparency and aid interpretation of AI systems.[8] Nevertheless, many studies treat explainability as a secondary analysis rather than integrating it into the model-development framework.

Despite substantial advances, important methodological limitations remain. Much research emphasizes optimization of predictive performance, whereas transparent and generalizable evaluation across settings remains limited. Few studies have examined behavioral and symptom-based features specifically for lung cancer status classification. Explainable AI has been applied to prediction, diagnosis, treatment, and management across chronic diseases.[19]

These gaps are also relevant to precision screening and prevention. The present study draws on concepts from precision medicine and explainable AI. Precision medicine emphasizes combining multiple patient-specific factors to estimate individualized risk, whereas explainable AI emphasizes transparent and interpretable processes for high-performing predictive systems.[18] Together, these perspectives motivate predictive frameworks that balance performance and interpretability.

The literature therefore suggests a need for integrated,

Table 1. Comparative analysis of previous studies and the present study

Study	Methodology	Strengths	Limitations	Key findings
Breiman (2001)[11]	Random forest	Strong predictive performance; reduced risk of overfitting	Limited interpretability	Ensemble methods can improve predictive performance
Chen & Guestrin (2016)[13]	XGBoost	High predictive performance and computational efficiency	Limited intrinsic interpretability	Improved performance on structured datasets
Ke *et al*. (2017)[12]	LightGBM	Fast and computationally efficient training	Moderate interpretability	Performance comparable to or better than conventional boosting methods
Esteva *et al*. (2019)[6]	Deep learning in healthcare	Handles complex data types	Requires large datasets; limited transparency	AI can improve diagnostic performance
Topol (2019)[8]	AI in precision medicine	Emphasizes clinical integration	Limited implementation frameworks	Explainability is important for adoption
Kourou *et al*. (2015)[7]	Review of machine learning for cancer prediction	Comprehensive evaluation of algorithms	No unified explainable framework	Machine learning can improve cancer-prediction performance
Present study	Explainable Precision Screening Framework (stacking ensemble: XGBoost + LightGBM + random forest; separate SVM SHAP analysis)	Combines ensemble classification with separate SVM-based SHAP interpretation	Requires external validation in clinically verified multicenter cohorts	Combines ensemble classification and exploratory SVM feature attribution in one framework

AI, artificial intelligence; SHAP, Shapley additive explanations; SVM, support vector machine.

explainable ensemble models that combine predictive performance with interpretable feature contributions based on behavioral and symptom variables. Rather than treating prediction and interpretability as separate objectives, a combined framework may support more transparent model evaluation and future validation.

To address this gap, this exploratory study proposes the EPSF, which integrates XGBoost, LightGBM, and random forest in a stacking ensemble, with a separate SHAP analysis of the standalone SVM model. The framework was designed to evaluate predictive performance while identifying feature patterns for subsequent external validation. Table 1 summarizes the methodological context and contributions of the present study.[6-8,11-13]

Materials and methods

Study design

The present study evaluates the EPSF for lung cancer status classification using ML, ensemble learning, and XAI. Rather than evaluating classification performance alone, the framework assesses both predictive performance and model interpretability within the analyzed dataset.

Data source and acquisition

The Lung Cancer Prediction Dataset hosted on Kaggle was used as the data source and was released under the Apache License 2.0.[20] The dataset contains de-identified, survey-based information from 309 records, including demographic characteristics, behavioral factors, symptoms, and a binary lung cancer label. The dataset was downloaded from its public repository and imported into Python for preprocessing and analysis.

Sample inclusion and exclusion criteria

Inclusion criteria

Records were included if they:

- had complete data for all predictor and outcome variables;
- had a definitive source-dataset lung cancer label of "YES" or "NO"; and
- had complete demographic, behavioral, and symptom-related variables.

Exclusion criteria

- Records with missing data for any study variable were excluded.
- Records with an unknown or ambiguous lung cancer label were excluded.

No records were excluded because all 309 observations had complete data for the analyzed variables, and no exact duplicate records were identified during data preprocessing. Participant selection and dataset processing are summarized in Figure 1.

Outcome definition

The study outcome was the source dataset's binary lung cancer label (YES/NO). No independent clinical verification of these

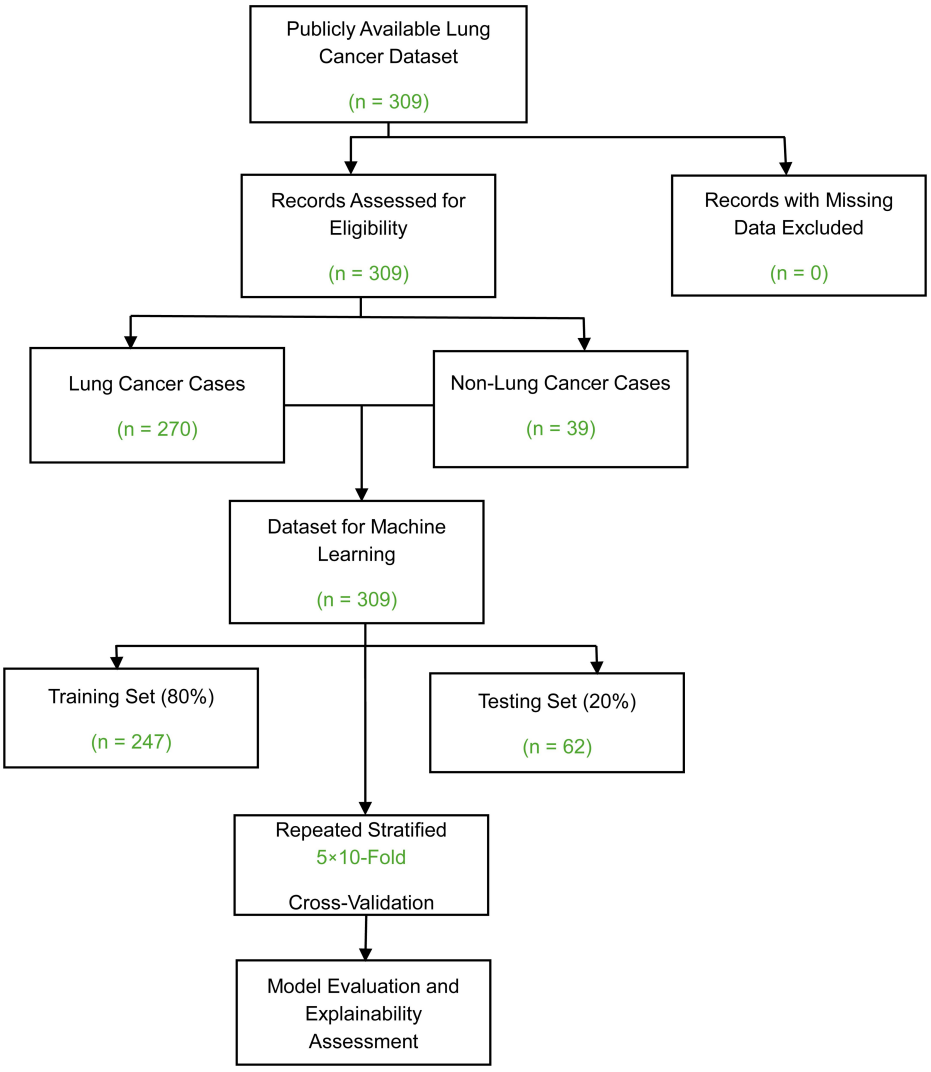

Fig. 1. Participant selection and dataset processing flow diagram. The participant-selection and data-processing flow diagram (Fig. 1) was prepared with reference to STROBE.[21] All 309 records in the public dataset were included because no missing values or exact duplicate records were identified. According to the source dataset labels, 270 records were labeled as lung cancer and 39 as non-cancer; the source dataset did not report the clinical reference standard used to establish these labels. The data were split into 80% for model development and 20% for hold-out testing. Repeated stratified cross-validation was conducted within the training data. STROBE, Strengthening the Reporting of Observational Studies in Epidemiology.

labels was available because this was a secondary analysis of a public dataset.

Definition of the control group

The control group comprised records labeled "NO" for lung cancer in the source dataset and served as the non-cancer reference group for model development and evaluation.

Reporting guideline considerations

This retrospective cross-sectional secondary-data study is reported with reference to the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement.[21] The dataset included demographic, behavioral, symptom, and binary lung cancer outcome variables. Although the source dataset included a lung cancer label, it did not report the clinical reference standard used to establish that label (e.g., histopathology, imaging, or physician diagnosis). Therefore, the study cannot fully comply with the Standards for Reporting Diagnostic Accuracy Studies (STARD) because the required clinical reference-standard information was unavailable.[22] All analyses relied on the source dataset labels without independent clinical verification.

Model development and validation workflow

Figure 2 summarizes the model-development, training, and evaluation workflow. The complete dataset (n = 309) was

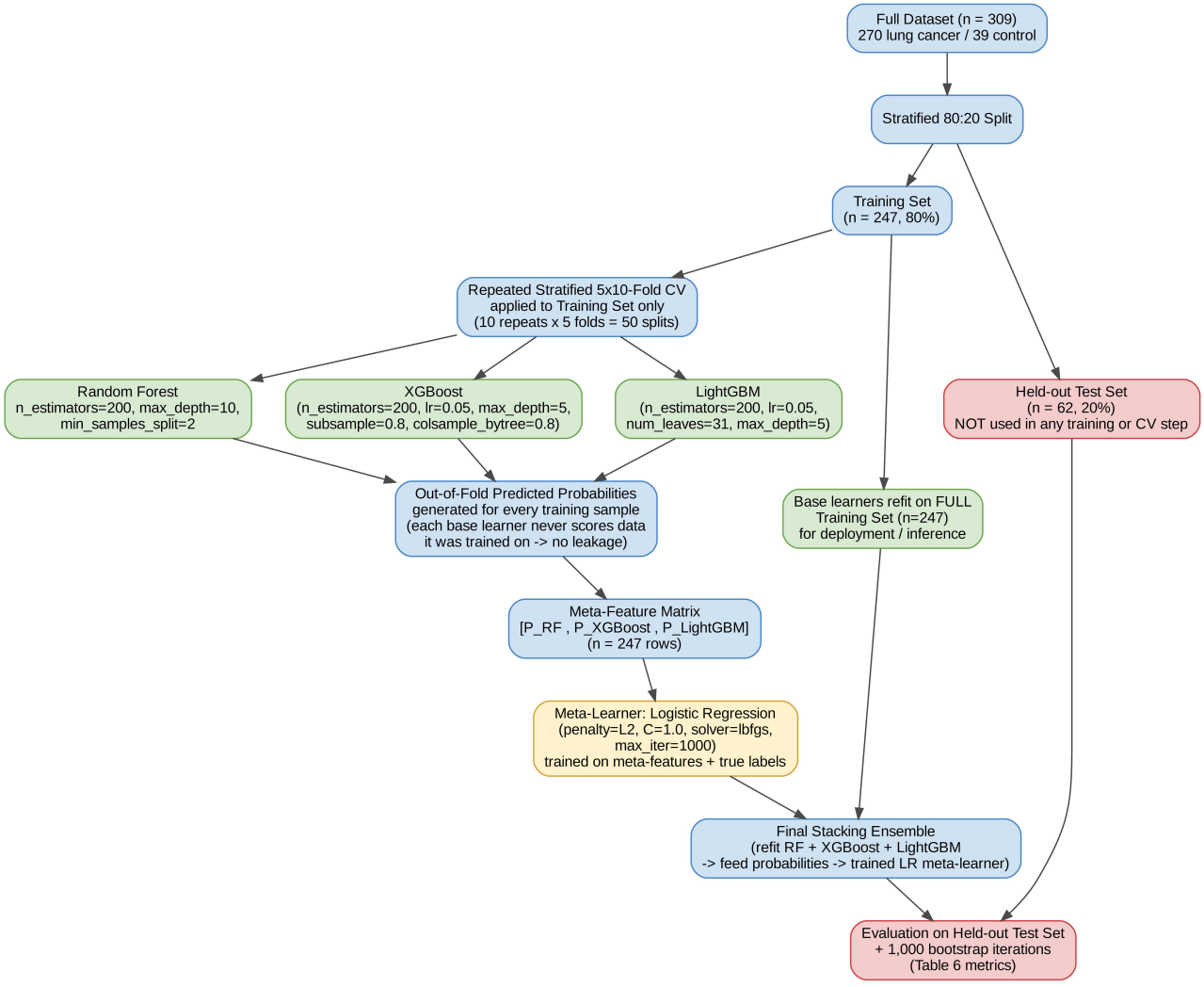

Fig. 2. Development and validation workflow of the Explainable Precision Screening Framework (EPSF). The dataset was split into a stratified training set (80%) and an independent hold-out test set (20%). Repeated stratified five-fold cross-validation was conducted within the training set to generate out-of-fold predictions from the three base learners (random forest, XGBoost, and LightGBM). These predictions were used to train the logistic-regression meta-learner. The base learners were then retrained on the full training set, and predictions for the hold-out test set were combined by the meta-learner. Model performance was evaluated using bootstrap 95% confidence intervals. CV, cross-validation; LightGBM, Light Gradient-Boosting Machine; LR, logistic regression; RF, random forest; XGBoost, Extreme Gradient Boosting;

stratified into a training set (80%; n = 247) and an independent hold-out test set (20%; n = 62) while maintaining the source class distribution. Model development used only the training data; the hold-out test set was reserved for final evaluation. Within the training set, repeated stratified five-fold cross-validation (10 repeats; 50 resampling iterations) was used to train the three base learners (random forest, XGBoost, and LightGBM). Out-of-fold predicted probabilities from cross-validation were used to construct the meta-feature matrix for training the logistic-regression meta-learner. After cross-validation, the three base learners were retrained on the full training dataset and used to generate predictions for the hold-out test set, which were combined by the trained meta-learner to produce the final ensemble predictions. Model performance was evaluated on the hold-out test set using accuracy, sensitivity, specificity, precision, negative predictive value, F1 score, ROC-AUC, PR-AUC, and Brier score, with 1,000 bootstrap iterations used to estimate 95% confidence intervals (CIs).

Data leakage prevention

To minimize the risk of data leakage, the independent hold-out test set was created before model development and used only for final evaluation. Within each cross-validation iteration, StandardScaler was fitted only on the corresponding training fold and then applied unchanged to its validation fold. After cross-validation, Standard-Scaler was fitted on the entire training set and then applied to the

hold-out test set. Repeated stratified five-fold cross-validation with 10 repeats and any hyperparameter tuning were restricted to the training data. No feature-selection or oversampling method, including the Synthetic Minority Over-sampling Technique, was used.

Variable coding

Table 2 summarizes the variable-coding scheme used during preprocessing.

Table 2. Variables coding scheme

Variable	Coding
Gender	Male = 1; female = 0
Age	Continuous variable (years)
Smoking	Yes = 1; no = 0
Yellow fingers	Yes = 1; no = 0
Anxiety	Yes = 1; no = 0
Peer pressure	Yes = 1; no = 0
Chronic disease	Yes = 1; no = 0
Fatigue	Yes = 1; no = 0
Allergy	Yes = 1; no = 0
Wheezing	Yes = 1; no = 0
Alcohol consumption	Yes = 1; no = 0
Coughing	Yes = 1; no = 0
Shortness of breath	Yes = 1; no = 0
Swallowing difficulty	Yes = 1; no = 0
Chest pain	Yes = 1; no = 0
Lung cancer	Yes = 1; no = 0

In the source dataset, most binary variables were encoded as 1 and 2. During preprocessing, these values were recoded as 0 and 1 for machine-learning analysis. Pairwise associations between binary predictors and the lung cancer label were summarized using Pearson correlations (equivalent to phi coefficients for binary variables), whereas age was summarized using the point-biserial correlation. The association between smoking status and yellow fingers was evaluated using a 2 × 2 contingency table, the phi coefficient, and Pearson's chi-square test.

Software environment and package versions

Analyses were performed using the following software:

- Python 3.11
- NumPy 1.26
- pandas 2.2
- scikit-learn 1.5
- XGBoost 2.1
- LightGBM 4.5
- SHAP 0.46
- Matplotlib 3.9
- StandardScaler from the scikit-learn preprocessing module

These software versions are reported to support reproducibility.

Hyperparameter settings

The following hyperparameters were used for model development:

- Random forest
- n_estimators = 200
- max_depth = 10
- min_samples_split = 2
- random_state = 42
- XGBoost
- n_estimators = 200
- learning_rate = 0.05
- max_depth = 5
- subsample = 0.8
- colsample_bytree = 0.8
- random_state = 42
- LightGBM
- n_estimators = 200
- learning_rate = 0.05
- num_leaves = 31
- max_depth = 5
- random_state = 42
- Meta-learner (logistic regression)
- penalty = L2
- C = 1.0
- solver = lbfgs
- max_iter = 1000

Standalone logistic regression used L2 regularization with C = 1.0, the lbfgs solver, and max_iter = 1000. The standalone SVM used a radial basis function kernel with C = 1.0, gamma = "scale", probability = True, and class_weight = None. These settings were prespecified based on preliminary tuning trials and guidance from the machine-learning literature.

SHAP analysis

For exploratory feature attribution, SHAP analysis was applied separately to the standalone SVM model using the model-development data (n = 247). A model-agnostic SHAP approach was applied to predicted probabilities for the lung cancer class. Mean absolute SHAP values were calculated across the analyzed observations and used to rank feature contributions; relative importance was calculated from each feature's mean absolute SHAP value relative to the sum across all features.

Sample size justification

No a priori power calculation was performed because the study was exploratory and based on secondary data. The analysis included all 309 complete records in the publicly available dataset and used repeated stratified five-fold cross-validation with 10 repeats for model evaluation.

Predictor handling and model-development considerations

Age, gender, smoking status, chronic illness status, and alcohol use were considered candidate predictors. The following strategies were used during model development:

1. All available clinically relevant variables were included in model development.
2. Multivariable models were used to evaluate predictors simultaneously.
3. Training and test sets were created using stratified sampling.
4. Repeated stratified five-fold cross-validation with 10 repeats

was used to reduce sampling variability.

5. A separate SHAP analysis of the standalone SVM model was used to characterize the contribution of each variable to its predictions.

The study focused on prediction rather than causal inference; therefore, causal-effect adjustment methods such as propensity-score matching were not applied.

Results

Baseline characteristics

Table 3 summarizes characteristics of the 309 records, including 270 lung cancer-labeled records and 39 non-cancer-labeled records. Among lung cancer-labeled records, the proportions with yellow fingers, coughing, wheezing, shortness of breath, alcohol consumption, and swallowing difficulty were 60.4%, 62.6%, 60.4%, 65.2%, 61.1%, and 51.9%, respectively. Mean age was 62.95 ± 7.97 years in lung cancer-labeled records and 60.74 ± 9.63 years in non-cancer-labeled records. Descriptive differences between the two groups were observed before model development.

Dataset composition

The dataset comprised 309 records: 270 (87.4%) labeled as lung cancer and 39 (12.6%) labeled as non-cancer. Fifteen demographic, behavioral, and symptom-related variables were evaluated.

Correlation analysis

Since the predictors and label are both binary outcomes. Hence Pearson's r is equivalent to the *phi* coefficient:

$$r_{\phi} = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

Where (*a*) and (*b*) are the predictor-positive and predictor-negative counts for the group having lung-cancer label. (*c*) and (*d*) are the respective counts of non-cancer group.

The information displayed in Table 4 and illustrated in Figure 3 presents the unadjusted pairwise correlations between the predictor variables and the lung cancer response variable in the original data set. Among the surveyed binary variables, the strongest and most positive correlation with the dependent variable was observed for allergy (r=0.328), followed by alcohol consumption (r=0.289) and swallowing difficulties (r=0.260). Similarly, wheezing (r=0.249) and coughing (r=0.249) were two other variables with almost identical correlations with the dependent variable, ranking fourth and fifth, respectively. Chest pain (r=0.190), peer pressure (r=0.186), and yellow fingers (r=0.181) were also positively correlated with the dependent variable but less strongly than the variables discussed above. Smoking (r=0.058) and breathlessness (r=0.061) were two variables for which the obtained correlation values with the dependent variable were the lowest.

These correlations indicate descriptive, unmoderated relationships among the imbalanced public dataset. These correlations do not approximate independent effects, establish causation, show clinical risk, or suggest that any variable is more important than others in multivariable prediction models.

Understanding Smoking and Yellow Fingers: Based on Figure 3, both smoking and yellow fingers are positively associated with the lung cancer label, but yellow fingers appear to have a stronger association. While yellow finger discoloration may imply tobacco exposure, this variable does not quantitatively measure the duration or intensity of smoking. In the same way, the binary measure of smoking does not capture the frequency or the cumulative amount of smoking. In the given data, yellow fingers

Table 3. Baseline characteristics of records according to source-dataset lung cancer status

Variable	Overall (n = 309)	Lung cancer label (n = 270)	Non-cancer label (n = 39)
Age (years), mean ± SD	62.67 ± 8.23	62.95 ± 7.97	60.74 ± 9.63
Gender			
Female, n (%)	147 (47.6%)	125 (46.3%)	22 (56.4%)
Male, n (%)	162 (52.4%)	145 (53.7%)	17 (43.6%)
Smoking	174 (56.3%)	155 (57.4%)	19 (48.7%)
Yellow fingers	176 (57.0%)	163 (60.4%)	13 (33.3%)
Anxiety	154 (49.8%)	142 (52.6%)	12 (30.8%)
Peer pressure	155 (50.2%)	145 (53.7%)	10 (25.6%)
Chronic disease	156 (50.5%)	142 (52.6%)	14 (35.9%)
Fatigue	208 (67.3%)	189 (70.0%)	19 (48.7%)
Allergy	172 (55.7%)	167 (61.9%)	5 (12.8%)
Wheezing	172 (55.7%)	163 (60.4%)	9 (23.1%)
Alcohol consumption	172 (55.7%)	165 (61.1%)	7 (17.9%)
Coughing	179 (57.9%)	169 (62.6%)	10 (25.6%)
Shortness of breath	198 (64.1%)	176 (65.2%)	22 (56.4%)
Swallowing difficulty	145 (46.9%)	140 (51.9%)	5 (12.8%)
Chest pain	172 (55.7%)	160 (59.3%)	12 (30.8%)

Continuous data are presented as mean ± standard deviation (SD); categorical data are presented as *n* (%). Percentages were calculated within each group. For binary variables, counts represent positive responses (coded as 2 in the source dataset).

Table 4. Pairwise correlations of predictors with the source-dataset lung-cancer label

Variable	Correlation with lung cancer label (r)
Allergy	0.328
Alcohol consumption	0.289
Swallowing difficulty	0.26
Wheezing	0.249
Coughing	0.249
Chest pain	0.19
Peer pressure	0.186
Yellow fingers	0.181
Fatigue	0.151
Anxiety	0.145
Chronic disease	0.111
Age	0.089
Gender (male)	0.067
Shortness of breath	0.061
Smoking	0.058

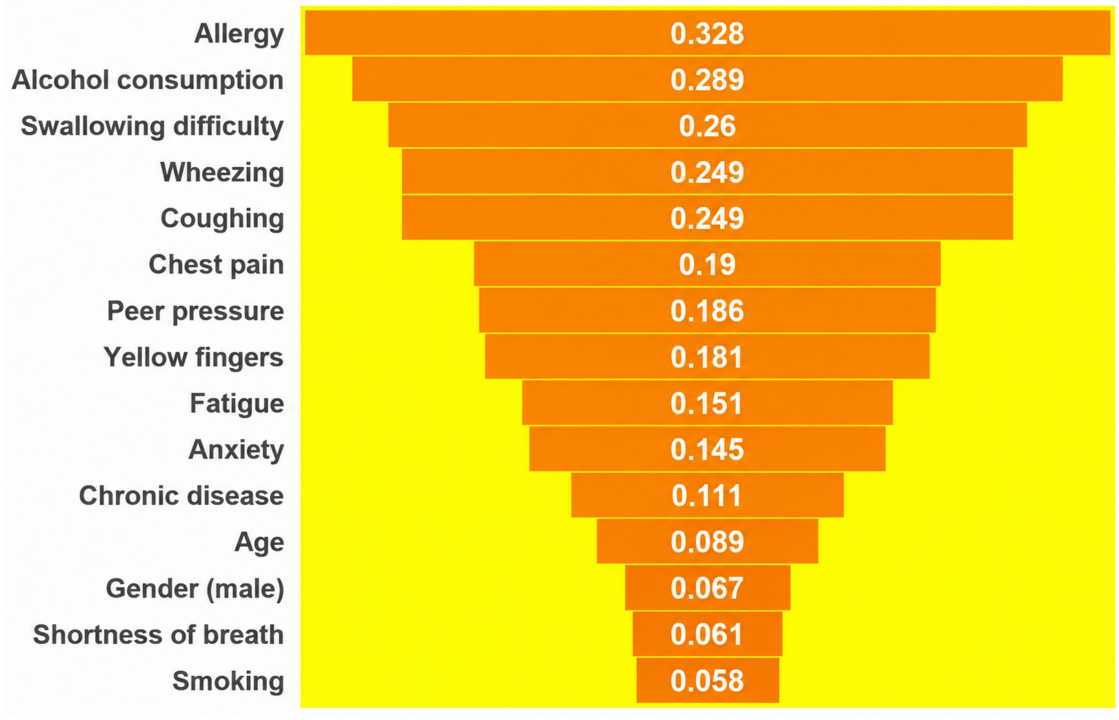

Fig. 3. Pairwise correlations of demographic, behavioral, and symptom variables with the source-dataset lung-cancer label.

appear to have a stronger pairwise correlation with the lung cancer label than the binary smoking variable. However, a direct 2 × 2 analysis of smoking status and yellow finger ascertained that there is practically no relationship between the two variables ($\varphi = -0.015$; $\chi^2(1) = 0.066$, $p = 0.798$). Thus, the stronger relationship between yellow fingers and the lung cancer label is not due to a strong association between smoking and yellow fingers in the given data. While the method used in this analysis is based on pairwise correlations, correlation does not imply an independent impact or causation.

Machine-learning model evaluation

Table 5 summarizes the validation strategy for all 309 records (270 lung cancer-labeled and 39 non-cancer-labeled). The data were split into training (80%; n = 247) and hold-out test (20%; n = 62) sets using stratified sampling. Repeated stratified five-fold cross-validation with 10 repeats was conducted within the training set, and 1,000 bootstrap iterations were used to estimate 95% CIs for hold-out performance metrics. Class imbalance was considered when interpreting model performance.

Table 5. Robustness and validation strategy

Validation component	Implementation
Dataset size	309 records
Positive records	270
Negative records	39
Test set	20% hold-out (n = 62)
Training set	80% (n = 247)
Cross-validation	Repeated stratified five-fold (10 repeats)
Bootstrap validation	1,000 iterations
Class imbalance monitoring	Yes
Performance reporting	Point estimates with 95% CIs
External validation	Not performed; recommended for future studies

CI, confidence interval.

Comparative model performance

As displayed in Table 6, all models showed high apparent performance for classifying the source dataset's lung cancer label. Logistic regression achieved an accuracy of 0.90 (95% CI: 0.82–0.97), an F1-score of 0.94 (95% CI: 0.90–0.98), and an ROC-AUC of 0.95 (95% CI: 0.88–0.99). Random forest and XGBoost yielded an accuracy of 0.92 (95% CI: 0.85–0.98 for random forest and 0.84–0.98 for the latter) and ROC-AUC of 0.95 (95% CI: 0.88–0.99 for the former and 0.89–0.99 for the latter). The stacking ensemble achieved the same accuracy point estimate as random forest and XGBoost. The stacking ensemble achieved an accuracy of 0.92 (95% CI: 0.85–0.98), precision of 0.96 (95% CI: 0.89–1.00), recall (sensitivity) of 0.94 (95% CI: 0.88–1.00), an F1 score of 0.95 (95% CI: 0.91–0.99), and ROC-AUC of 0.95 (95% CI: 0.89–0.99). Its PR-AUC was 0.993 (95% CI: 0.981–0.999), and its Brier score was 0.074 (95% CI: 0.037–0.121). Corresponding PR-AUC and Brier score estimates for all models are provided in Supplementary File 1. Figure 4 summarizes model performance. The bootstrap confidence intervals reflect uncertainty in performance estimates for this dataset.

To broaden the internal assessment, performance reporting included specificity and negative predictive value (NPV) in addition to accuracy, precision, recall, F1 score, and ROC-AUC. Model robustness was further assessed using repeated stratified cross-validation together with a strictly independent hold-out test set to reduce optimism and overfitting. External validation was outside the scope of this study; future work should evaluate the EPSF in independent, clinically verified multicenter cohorts to assess generalizability across populations.

Examination of the confusion matrix and performance evaluation

In addition to the metrics reported in Table 6, Table 7 presents true-positive, false-negative, true-negative, and false-positive counts together with balanced accuracy and Matthews correlation coefficient for each model. These measures provide additional information under class imbalance: balanced accuracy weights sensitivity and specificity equally, whereas Matthews correlation coefficient summarizes overall binary classification performance. Consistent with Table 6, the stacking ensemble performed similarly to Random Forest and XGBoost. Detailed calculations of the performance metrics are provided in Supplementary File 1.

Explainable feature analysis

Table 8 summarizes mean absolute SHAP values from the standalone SVM analysis. Smoking had the largest mean absolute SHAP value (0.284; 22.8%), followed by yellow fingers (0.241; 19.3%), coughing (0.176; 14.1%), chest pain (0.139; 11.2%), wheezing (0.112; 9.0%), shortness of breath (0.097; 7.8%), and age (0.076; 6.1%). These values reflect model-based feature contributions on the model-development data and do not establish statistical significance, causal effects, or clinical risk.

Discussion

This study developed and evaluated an ensemble machine-learning framework for classifying source-dataset lung cancer status using demographic, behavioral, and symptom variables. The EPSF evaluates a stacking ensemble comprising XGBoost, LightGBM, and random forest, with SHAP used separately to interpret the standalone SVM model.23-25 On the hold-out test set, the stacking ensemble achieved accuracy of 92%, precision of 96%, recall of 94%, an F1 score of 95%, and ROC-AUC = 0.951. These findings show that the stacking approach achieved performance comparable to the highest-performing individual models within the current dataset.

Prior studies have reported strong performance of boosting and ensemble methods in healthcare prediction tasks. Chen and Guestrin described XGBoost as an efficient approach for modeling nonlinear relationships and feature interactions[13]; Ke *et al*.[12] described LightGBM as a computationally efficient gradient-boosting method for structured data. Likewise, Kourou *et al*.[7] reviewed machine-learning methods for cancer prediction and highlighted the potential of ensemble approaches. In the present study, the stacking architecture achieved performance comparable to the highest-performing individual models in this dataset.[23-25]

The unmodified correlation analysis and the independent SVM SHAP analysis focus on different areas of the data and should not be confused with one another. In the correlation analysis, allergy, alcohol consumption, and difficulty swallowing were the most strongly correlated variables with the lung cancer label in the source dataset. On the other hand, in the independent SVM analysis, the largest SHAP values were recorded for smoking,

Table 6. Comparative performance metrics on the independent hold-out test set (n = 62) with bootstrap 95% confidence intervals

Model	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	Precision (95% CI)	NPV (95% CI)	F1 score (95% CI)	ROC-AUC (95% CI)
Logistic regression	0.90 (0.82–0.97)	0.94 (0.87–1.00)	0.62 (0.25–1.00)	0.94 (0.88–1.00)	0.62 (0.27–1.00)	0.94 (0.90–0.98)	0.95 (0.88–0.99)
SVM	0.85 (0.77–0.94)	0.94 (0.87–1.00)	0.25 (0.00–0.60)	0.89 (0.81–0.97)	0.40 (0.00–1.00)	0.92 (0.87–0.97)	0.93 (0.86–0.99)
Random forest	0.92 (0.85–0.98)	0.94 (0.88–1.00)	0.75 (0.40–1.00)	0.96 (0.91–1.00)	0.67 (0.33–1.00)	0.95 (0.91–0.99)	0.95 (0.88–0.99)
LightGBM	0.90 (0.84–0.97)	0.94 (0.88–1.00)	0.62 (0.25–1.00)	0.94 (0.88–1.00)	0.62 (0.25–1.00)	0.94 (0.90–0.98)	0.93 (0.85–0.99)
XGBoost	0.92 (0.84–0.98)	0.94 (0.88–1.00)	0.75 (0.43–1.00)	0.96 (0.91–1.00)	0.67 (0.33–1.00)	0.95 (0.91–0.99)	0.95 (0.89–0.99)
Stacking ensemble	0.92 (0.85–0.98)	0.94 (0.88–1.00)	0.75 (0.40–1.00)	0.96 (0.89–1.00)	0.67 (0.33–1.00)	0.95 (0.91–0.99)	0.95 (0.89–0.99)

CI, confidence interval; NPV, negative predictive value; ROC-AUC, area under the receiver operating characteristic curve; SVM, support vector machine.

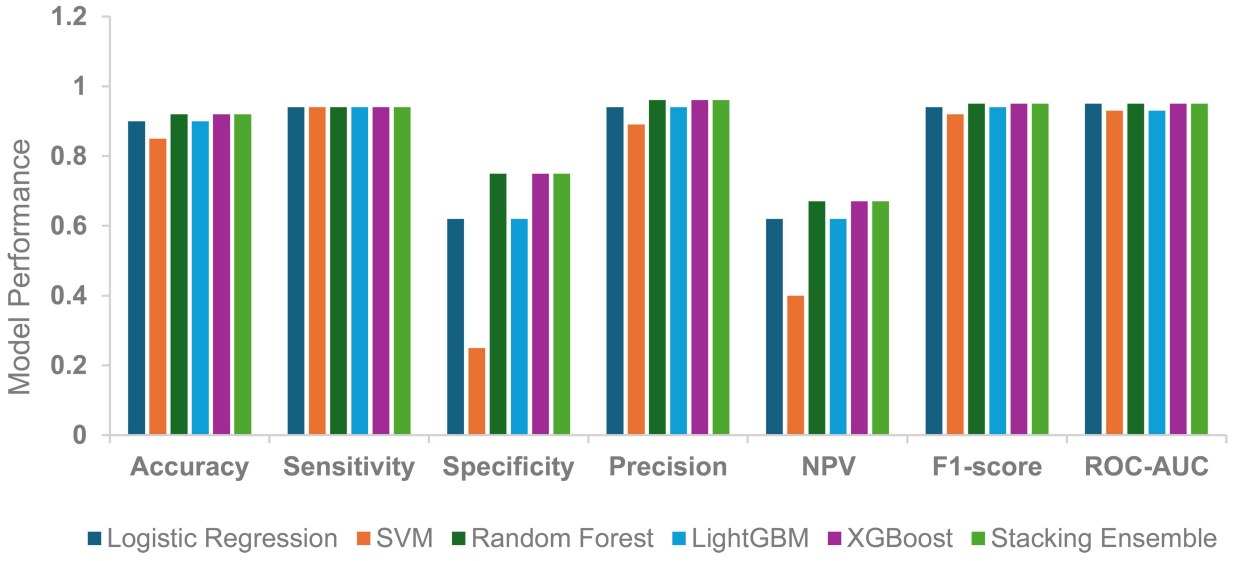

Fig. 4. Performance of machine-learning algorithms. CI, confidence interval; NPV, negative predictive value; ROC-AUC, area under the receiver operating characteristic curve; SVM, support vector machine.

yellow fingers, coughing, chest pain, wheezing, shortness of breath, and age. While the correlations indicate marginal dependence between the variables and the lung cancer label, the SHAP values indicate the relative importance of each variable, accounting for the effects of other variables and the trained SVM. Thus, the discrepancy between the two rankings is anticipated and does not indicate that these variables are independent clinical risk factors or clinically validated biomarkers.

Model generalizability and potential biases

The dataset showed marked class imbalance (270 lung cancer-labeled records [87.4%] versus 39 non-cancer-labeled records [12.6%]). This imbalance may bias model development toward the majority class and inflate overall accuracy. Stratified partitioning and repeated stratified cross-validation,[26] and complementary performance metrics (sensitivity, specificity, precision, NPV, F1 score, and ROC-AUC) were used to characterize performance under class imbalance; however, larger and more balanced datasets are needed for robust evaluation. Second, a single public dataset cannot represent demographic, geographic, ethnic, socioeconomic, and healthcare-system variation. The models

Table 7. Classification outcomes on the independent hold-out test set (*n* = 62)

Model	TP	FN	TN	FP	(N = 62)	Accuracy	Specificity	Balanced accuracy	Precision	MCC
Logistic regression	51	3	5	3	62	90.32%	62.50%	78.47%	94.44%	0.569
SVM	51	3	2	6	62	85.48%	25.00%	59.72%	89.47%	0.239
Random forest	51	3	6	2	62	91.94%	75.00%	84.72%	96.23%	0.661
LightGBM	51	3	5	3	62	90.32%	62.50%	78.47%	94.44%	0.569
XGBoost	51	3	6	2	62	91.94%	75.00%	84.72%	96.23%	0.661
Stacking ensemble	51	3	6	2	62	91.94%	75.00%	84.72%	96.23%	0.661

FN, false negative; FP, false positive; MCC, Matthews correlation coefficient; SVM, support vector machine; TN, true negative; TP, true positive.

Table 8. SVM predictive-feature ranking based on mean absolute SHAP values

Rank	Variable	Mean absolute SHAP value	Relative importance (%)
1	Smoking	0.284	22.8
2	Yellow fingers	0.241	19.3
3	Coughing	0.176	14.1
4	Chest pain	0.139	11.2
5	Wheezing	0.112	9.0
6	Shortness of breath	0.097	7.8
7	Age	0.076	6.1
8	Fatigue	0.049	3.9
9	Swallowing difficulty	0.034	2.7
10	Chronic disease	0.021	1.7
11	Alcohol consumption	0.011	0.9
12	Peer pressure	0.005	0.4
13	Anxiety	0.003	0.2
14	Gender	0.002	0.2
15	Allergy	0.001	0.1

SHAP, Shapley additive explanations; SVM, support vector machine.

should therefore be considered preliminary pending external validation in independent, diverse multicenter cohorts.[27]

The proposed EPSF should be interpreted as an exploratory classification model, not as a clinical decision-support or diagnostic tool. The current findings may be used to guide further research, but external validation in clinically verified cohorts is required before considering any clinical application. Because the model was developed using demographic, behavioral, and symptom-based variables from a single public dataset, its performance should not be generalized to asymptomatic screening populations, pediatric populations, or other clinical populations.

Although the framework showed strong internal performance, the findings remain exploratory. Independent external validation using geographically diverse, clinically verified samples is needed to assess reproducibility, calibration, and generalizability before any clinical implementation. The current findings therefore do not demonstrate clinical utility.

Theoretical standpoint

The findings are consistent with the precision-medicine principle that individualized risk assessment can integrate multiple patient-specific factors rather than relying on a single risk factor.8 The study also illustrates the explainable-AI principle that predictive models can be accompanied by transparent feature-attribution methods. Prior work has highlighted a potential trade-off between predictive performance and interpretability.18 The present framework combines ensemble learning with a separate SHAP-based interpretation of the standalone SVM model, although its clinical value remains to be established.

Study implications and future applications

Current lung cancer screening programs primarily use LDCT for eligible high-risk populations. Although effective, LDCT implementation may be limited by cost, access, false-positive findings, and radiation exposure.[3-5] The present study suggests a future research direction in which demographic, behavioral, and symptom-based variables are evaluated as complements to established screening approaches. However, the EPSF has not been externally validated and should not be implemented in electronic health records, telemedicine platforms, or primary-care screening programs on the basis of this study. Its potential value in resource-limited settings requires evaluation in clinically verified, representative cohorts.

Recent literature likewise emphasizes that healthcare AI requires a balance of predictive performance and explainability, clinical validation, regulatory consideration, and integration into clinical workflows.9 External validation and clinical assessment are therefore essential before explainable machine-learning frameworks are used for decision support.

Age-related findings

In this dataset, age had a smaller mean absolute SHAP value than several behavioral and symptom-related features. Age is an established lung cancer risk factor.[15] However, the observed SHAP ranking may reflect the composition and limited size of this dataset and should not be interpreted as evidence that age is less clinically important. Larger and more diverse cohorts are needed to determine whether this pattern is reproducible.

Table 9. STARD considerations

STARD requirement	Status in current study	Explanation
Clinical reference standard	Not reported by the source dataset	Diagnosis labels were present, but no information about the clinical reference standard was available in the original dataset. As a result, it was not possible to conduct independent clinical verification.
Blinding of diagnostic assessment	Not applicable	No human interpretation involved
Inter-rater reliability (kappa)	Not applicable	No multiple raters
Intra-rater reliability	Not applicable	No repeated human assessment
Sensitivity reported	Yes	Included
Specificity reported	Yes	Added
PPV reported	Yes	Included as Precision
NPV reported	Yes	Added
Independent test set	Yes	20% hold-out set
Repeated cross-validation	Yes	Repeated stratified 5×10-fold
External validation	No	Recommended for future work

NPV, negative predictive value; PPV, positive predictive value; STARD, Standards for Reporting Diagnostic Accuracy Studies.

Knowledge contribution

Although limited, this study contributes an exploratory example of combining ensemble learning with separate SHAP-based interpretation of a standalone SVM model. The analysis identified the variables that contributed most strongly to SVM predictions within one public dataset. These findings do not establish clinically actionable risk signatures or readiness for deployment; rather, they provide a basis for external validation and further methodological research.

Considerations related to STARD

Interpretation of this study is limited by incomplete applicability of the STARD guidelines.[22] The source dataset included a binary lung cancer label but did not report the clinical reference standard used to establish that label. Consequently, the reference standard could not be independently verified, and the findings should be interpreted as classification of source-dataset labels rather than diagnostic accuracy against a clinically validated standard.

To reduce bias in model evaluation, the data were partitioned into an independent hold-out test set and a training set subjected to repeated stratified cross-validation. Predictors and outcome labels were taken directly from the source dataset; no human raters performed diagnostic interpretation. Accordingly, blinding of diagnostic assessors, inter-rater reliability, intra-rater reliability, and Cohen's kappa were not applicable.

To improve reporting of classification performance, specificity and NPV were reported in addition to sensitivity, precision, F1 score, accuracy, and ROC-AUC. Future evaluations of the framework should use clinically verified multicenter datasets with documented reference standards and independent external validation to support fuller STARD reporting. Table 9 summarizes the STARD items applicable to the present study.

Study limitations

Several limitations should be considered. First, the EPSF underwent internal validation only, using an independent hold-out test set and repeated stratified five-fold cross-validation with 10 repeats. Because no independent external dataset was used, the generalizability of the proposed explainable precision screening framework to different clinical populations has not yet been established. Future studies are needed with geographically diverse and prospectively collected datasets to evaluate the model's reproducibility, calibration, and generalizability.

The dataset was markedly imbalanced, with 270 lung cancer-labeled records (87.4%) and 39 non-cancer-labeled records (12.6%). This imbalance may bias model development toward the majority class and inflate overall accuracy. Stratified train-test partitioning, repeated stratified cross-validation, and complementary metrics (sensitivity, specificity, precision, NPV, F1 score, and ROC-AUC) were used to characterize performance beyond overall accuracy. The limited number of non-cancer records may particularly affect estimates of specificity and calibration.

Future studies should use larger, more balanced cohorts and evaluate additional approaches to class imbalance, such as the Synthetic Minority Over-sampling Technique, adaptive class weighting, or cost-sensitive learning, together with independent external validation.[28]

Conclusions

This retrospective cross-sectional study developed a machine-learning framework (EPSF) for classifying source-dataset lung cancer labels using demographic, behavioral, and symptom-based variables from a publicly available dataset. The stacking ensemble showed strong internal discrimination comparable to the highest-performing individual models, and separate SHAP analysis of the standalone SVM model identified the features contributing most strongly to its predictions.

The framework remains exploratory because it was evaluated using a single retrospective public dataset with internal validation only. Independent external validation in large, clinically verified multicenter cohorts is required before clinical use. Because the source labels were not verified against a reported clinical reference standard, these findings reflect classification of the dataset labels rather than clinically validated diagnosis. Thus, the results can be seen only as preliminary evidence and should be followed by further research.

Supporting information

Supplementary material for this article is available at https://doi.org/10.14218/CSP.2026.00009.

Acknowledgments

The author thanks the School of Industrial Sciences and Technology at the University of Central Missouri for supporting this study.

Funding

No funding was received.

Conflict of interest

The author declares no conflicts of interest.

Author contributions

HS is the sole author of the manuscript.

Ethical statement

Because the study utilized a publicly accessible, de-identified dataset with no personally identifiable information and no direct human subject involvement, Institutional Review Board (IRB) approval and informed consent were not required. The study complied with ethical principles for secondary data analysis and the Declaration of Helsinki (as revised in 2024), as applicable to secondary analyses of publicly available de-identified data. No formal exemption number exists.

Data sharing statement

The dataset analyzed in this study is publicly available from Kaggle (Survey Lung Cancer; updated August 15, 2023; Apache License 2.0) at https://www.kaggle.com/datasets/awais8765/survey-lung-cancer and was accessed on June 1, 2026. The analytical code developed and used for this study is not publicly available.

References

[1] Siegel RL, Kratzer TB, Giaquinto AN, Sung H, Jemal A. Cancer statistics, 2025. CA Cancer J Clin 2025;75(1):10–45. DOI: 10.3322/caac.21871, PMID: 39817679.

[2] World Health Organization. Lung cancer. Geneva: World Health Organization; 2026. Available from: https://www.who.int/news-room/fact-sheets/detail/lung-cancer. Accessed July 24, 2026.

[3] de Koning HJ, van der Aalst CM, de Jong PA, Scholten ET, Nackaerts K, Heuvelmans MA, *et al*. Reduced Lung-Cancer Mortality with Volume CT Screening in a Randomized Trial. N Engl J Med 2020;382(6):503–513. DOI: 10.1056/NEJMoa1911793, PMID: 31995683.

[4] Jonas DE, Reuland DS, Reddy SM, Nagle M, Clark SD, Weber RP, *et al*. Screening for Lung Cancer With Low-Dose Computed Tomography: Updated Evidence Report and Systematic Review for the US Preventive Services Task Force. JAMA 2021;325(10):971–987. DOI: 10.1001/jama.2021.0377, PMID: 33687468.

[5] Sedani AE, Davis OC, Clifton SC, Campbell JE, Chou AF. Facilitators and Barriers to Implementation of Lung Cancer Screening: A Framework-Driven Systematic Review. J Natl Cancer Inst 2022;114(11):1449–1467. DOI: 10.1093/jnci/djac154, PMID: 35993616.

[6] Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, *et al*. A guide to deep learning in healthcare. Nat Med 2019;25(1):24–29. DOI: 10.1038/s41591-018-0316-z, PMID: 30617335.

[7] Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J 2015;13:8–17. DOI: 10.1016/j.csbj.2014.11.005, PMID: 25750696.

[8] Topol E. Deep medicine: how artificial intelligence can make healthcare human again. New York: Basic Books; 2019.

[9] Xie H, Jia Y, Liu S. Integration of artificial intelligence in clinical laboratory medicine: Advancements and challenges. Interdiscip Med 2024;2(3):e20230056. DOI: 10.1002/inmd.20230056.

[10] Makubhai SS, Pathak GR, Chandre PR. Predicting lung cancer risk using explainable artificial intelligence. Bull Electr Eng Inform 2024;13(2):1276–1285. DOI: 10.11591/eei.v13i2.6280.

[11] Breiman L. Random forests. Mach Learn 2001;45(1):5–32. DOI: 10.1023/A:1010933404324.

[12] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, *et al*. LightGBM: A highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst 2017;30:3146–3154.

[13] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016 Aug 13-17; San Francisco, CA, USA. New York: Association for Computing Machinery; 2016:785-794. DOI: 10.1145/2939672.2939785.

[14] Ganie SM, Dutta Pramanik PK, Zhao Z. Enhanced and Interpretable Prediction of Multiple Cancer Types Using a Stacking Ensemble Approach with SHAP Analysis. Bioengineering (Basel) 2025;12(5):472. DOI: 10.3390/bioengineering12050472, PMID: 40428091.

[15] Callender T, Imrie F, Cebere B, Pashayan N, Navani N, van der Schaar M, *et al*. Assessing eligibility for lung cancer screening using parsimonious ensemble machine learning models: A development and validation study. PLoS Med 2023;20(10):e1004287. DOI: 10.1371/journal.pmed.1004287, PMID: 37788223.

[16] Kumar S, Gota V. Logistic regression in cancer research: A narrative

review of the concept, analysis, and interpretation. Cancer Res Stat Treat 2023;6(4):573–578. DOI: 10.4103/crst.crst_293_23.

[17] Ayer T, Chhatwal J, Alagoz O, Kahn CE Jr, Woods RW, Burnside ES. Informatics in radiology: comparison of logistic regression and artificial neural network models in breast cancer risk estimation. Radiographics 2010;30(1):13–22. DOI: 10.1148/rg.301095057, PMID: 19901087.

[18] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 2017;30:4765–4774.

[19] Hoghooghi Esfahani H, Toyonaga S, Oyibo K. The application of explainable artificial intelligence in the prediction, diagnoses, treatment, and management of chronic diseases: A systematic review. Digit Health 2025;11:20552076251355669. DOI: 10.1177/20552076251355669, PMID: 41312145.

[20] Awais8765. Survey Lung Cancer [dataset on the Internet]. Kaggle. Updated 2023. Apache License 2.0. Available from: https://www.kaggle.com/datasets/awais8765/survey-lung-cancer. Accessed June 1, 2026.

[21] von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP; STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. PLoS Med 2007;4(10):e296. DOI: 10.1371/journal.pmed.0040296, PMID: 17941714.

[22] Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, *et al*. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ 2015;351:h5527. DOI: 10.1136/bmj.h5527, PMID: 26511519.

[23] Alonso E, Calle X, Gurrutxaga I, Beristain A. Survival Stacking Ensemble Model for Lung Cancer Risk Prediction. Stud Health Technol Inform 2024;321:155–159. DOI: 10.3233/SHTI241083, PMID: 39575799.

[24] Bhaskar PV, Veeramani R, Rao KB, Devi EL, Saripalle S, Dhumpati R. An interpretable stacking ensemble model AB-CBLC for high-accuracy lung cancer detection using clinical and behavioral data. In: 2025 8th International Conference on Computing Methodologies and Communication (ICCMC); 2025 Jul 23-25; Erode, India. Piscataway (NJ): IEEE; 2025:1580-1587. DOI: 10.1109/ICCMC65190.2025.11140773.

[25] Tu H, Zhao Y, Cui J, Lu W, Sun G, Xu X, *et al*. Improving Lung Cancer Risk Prediction Using Machine Learning: A Comparative Analysis of Stacking Models and Traditional Approaches. Cancers (Basel) 2025;17(10):1651. DOI: 10.3390/cancers17101651, PMID: 40427148.

[26] Yates LA, Aandahl Z, Richards SA, Brook BW. Cross-validation for model selection: A review with examples from ecology. Ecol Monogr 2023;93(1):e1557. DOI: 10.1002/ecm.1557.

[27] Steyerberg EW, Harrell FE Jr. Prediction models need appropriate internal, internal-external, and external validation. J Clin Epidemiol 2016;69:245–247. DOI: 10.1016/j.jclinepi.2015.04.005, PMID: 25981519.

[28] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. J Artif Intell Res 2002;16:321–357. DOI: 10.1613/jair.953.