Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871

Vol. 6, No. 4, Agustus 2025, Page. 2187-2206
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4925

Visual Interpretation of Machine Learning Models (Random Forest) for Lung Cancer Risk Classification Using Explainable Artificial Intelligence (SHAP & LIME)

Irwan Fathur Rosyid[*1], Himawan Pramaditya[2]

[1,2]Information System, Faculty of Information Technology, Universitas Merdeka Malang, Indonesia

Email: [1]irwanfathurrosyid2003@gmail.com

Received : Jun 20, 2025; Revised : Jul 14, 2025; Accepted : Jul 20, 2025; Published : Aug 19, 2025

Abstract

Lung cancer remains one of the most prevalent and burdensome cancers worldwide, with delayed diagnosis being a persistent challenge—particularly in Indonesia, where no national screening program currently exists. In this collaborative study, we aim to develop an interpretable machine learning model for classifying lung cancer risk levels using the Explainable Artificial Intelligence (XAI) approach. The CRISP-DM framework was applied, and the dataset underwent cleaning, feature selection, labeling, and transformation, resulting in 152 valid entries. Tree ensemble algorithms—XGBoost, Random Forest, and LightGBM—were used, with Random Forest achieving the best performance at 97.38% accuracy. SHAP and LIME methods were integrated to provide transparent visual interpretations. A web-based system was developed using Streamlit, incorporating these visualizations and automated narrative summaries generated by a language model to assist non-technical users. A simulated case based on a published pediatric lung cancer report was used to demonstrate its interpretability and illustrate its potential applicability in clinical workflows. The proposed system offers an interpretable and scalable solution for early lung cancer risk classification, which may enhance decision support in primary care and promote trust in AI-assisted diagnostics.

Keywords : *CRISP-DM, Explainable AI, Lung Cancer, Machine Learning, Risk Classification.*

1. INTRODUCTION

The increasing number of lung cancer patients has become one of the threats to public health. Lung cancer is a situation where abnormal cells grow uncontrollably in the lungs, forming tumors that cause breathing problems because they do not grow into healthy lung tissue [1]. Lung cancer is the most commonly diagnosed type of cancer in men worldwide, ranking first among all cancer cases in the male population [2]. Lung cancer patients are found in 37 countries, including Russia, China, Eastern Europe, the Middle East, and Southeast Asia [3]. Indonesia is part of the Southeast Asian region. According to data from the Global Cancer Observatory (GLOBOCAN) in 2020, the incidence of lung cancer in men in Indonesia was estimated at 19.4 per 100,000 people [4]. By 2022, the number of cases had increased to 21.3 per 100,000 people [5].

The increased number of these cases shows that risk factors are still common in society. One of the most significant risk factors is smoking [6]. Smokers have a tenfold higher risk due to inhaling tobacco smoke, which consists of 4,000 harmful chemical compounds [7]. Smoking is an integral part of Indonesian social and cultural life, as evidenced by the high number of active smokers, estimated at around 69 million people [4]. Many other factors can increase an individual's risk of lung cancer, including air pollution, passive smoking, chronic lung disease, genetic risk, gender, and age [3], [6], [8]-[11]. The relationships between these factors are complicated, which makes it difficult to identify at-risk groups in a correct and organized way.

Vol. 6, No. 4, Agustus 2025, Page. 2187-2206
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4925

The challenge of identifying high-risk groups is made more difficult by the complexity of these various risk factors in a healthcare system that is not yet fully capable of supporting early detection. In Indonesia, most lung cancer patients are only diagnosed when the disease has reached an advanced stage [12]. This significantly reduces life expectancy. Until now, there has been no national screening program for lung cancer that is organized in a systematic way [4]. The methods that are used most often for diagnosis still use imaging technologies such as CT scans, but these have limitations in terms of cost, access, and expertise [4], [13].

On the other hand, most of the information related to these risk factors is already documented in routine clinical data—either through electronic health records (EHR) or primary care records such as those at community health centers or outpatient clinics [14], [15], [16]. This data includes information such as age, smoking status, history of lung disease, and other factors that play a role in risk classification. Therefore, a solution is needed that can efficiently manage this clinical data and support more accurate early detection processes.

Artificial intelligence (AI) can be used to address these limitations. Artificial intelligence (AI) is defined as the process of developing computers or computational systems that can perform tasks typically requiring human intelligence [17]. One of the main branches of AI that is rapidly developing today is Machine Learning (ML). This method enables computers to learn from data and improve their performance automatically without explicit programming [18]. These methods can efficiently process clinical data and generate risk predictions [19].

However, ML models are generally 'black box' and difficult for medical professionals to understand [20]. Explainable AI (XAI) aims to bridge this gap by transforming models that were initially black boxes into glass box models (explanations that are understandable to humans) [21]. Previous studies have evaluated the effectiveness of ML algorithms in predicting lung cancer: Dritsas and Trigka (2022) used Random Forest and Rotation Forest [22]; Mamun et al. (2022) applied XGBoost, LightGBM, Bagging, and AdaBoost [23]; Sweet et al. (2024) compared XGBoost, SVM, and Logistic Regression [24]; and Pathan et al. (2024) developed a model using Gradient Boosting, RF, DT, and Logistic Regression [25]. While Pathan et al. applied the SHAP interpretation method as an XAI approach, their research has not primarily focused on interactive, case study-based interpretive visualization.

Most of these studies still focus on model performance without emphasizing interpretability based on case studies that are easily accessible to non-technical medical personnel. Additionally, few studies place this risk classification system within the context of its application as part of a national screening policy framework, particularly in developing countries like Indonesia. This highlights a gap that needs to be addressed through research that not only prioritizes model accuracy and transparency but also ensures the system's utility at the level of systematic early detection.

This study aims to address these concerns by developing a lung cancer risk prediction model using machine learning algorithms and applying Explainable Artificial Intelligence (XAI) methods to provide visual explanations of the predictions in a case-based context. The research involved building a classification model trained on open-source data and applying XAI interpretation methods via a graphical user interface (GUI) that is easier for non-technical users to understand.

2. METHOD

This study employs the CRISP-DM (Cross-Industry Standard Process for Data Mining) approach as its methodological framework. CRISP-DM is a well-established process model that has been widely accepted in cross-industry data mining practices due to its flexibility and independence from specific algorithms [26], [27]. Figure 1 shows the methodological flow used in this study based on the CRISP-DM framework.

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 4, Agustus 2025, Page. 2187-2206
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4925

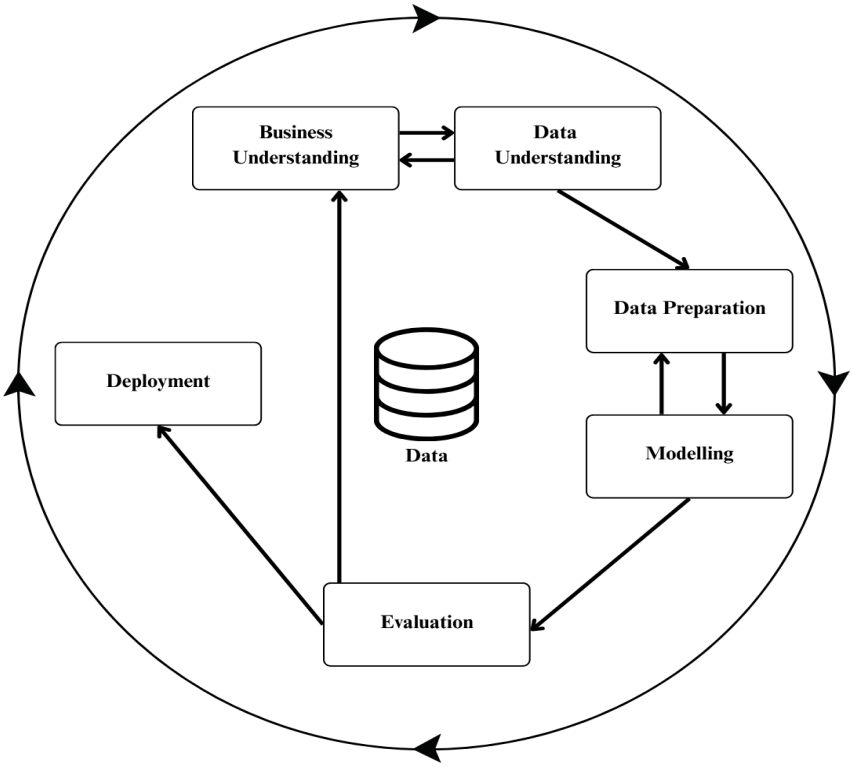


Figure 1. CRISP-DM Framework

This framework comprises six main stages: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment. Each stage will be explained according to its implementation in this study.

2.1. Business Understanding

The first stage aims to understand the problem and objectives of this study. Understanding the context of the problem is crucial in assessing the limitations of conventional approaches and designing solutions that are relevant to user needs [30]. The main problem in this research is the high risk of lung cancer, which is often detected late, and the lack of a prediction system that can be interpreted by medical personnel. Therefore, this study focuses on developing a machine learning model that is not only accurate but also explainable (interpretable) through the Explainable AI (XAI) approach. Next, business objectives are determined and then translated into structured data mining objectives [27]. The primary objectives of this study are to build a lung cancer risk classification model, integrate XAI interpretation methods, and present interpretation results through a web-based prototype interface to enhance clinical understanding of prediction outcomes.

2.2. Data Understanding

This stage involves the initial exploration of the data. This process involves collecting data from various sources, describing the data, conducting visual and statistical explorations, and assessing data quality [27]. Data structure and quality are important foundations before entering the data transformation stage [28]. The dataset used in this study was obtained from Kaggle [29]. The Lung Cancer Dataset used consists of 25 features and 1000 entries representing lung cancer patients with different risk levels, such as *low*, *medium*, and *high.* These features represent risk factors and symptoms related to lung cancer [6]. Features representing risk factors are *Age, Gender, Air Pollution, Alcohol use, Dust Allergy, Occupational Hazards, Genetic Risk, Chronic Lung Disease, Balanced Diet, Obesity, Smoking,* and *Passive Smoker*. Meanwhile, the features representing symptoms are *Chest Pain, Coughing of Blood, Fatigue, Weight Loss,*

Vol. 6, No. 4, Agustus 2025, Page. 2187-2206
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4925

Shortness of Breath, Wheezing, Swallowing Difficulty, Clubbing of Finger Nails, Frequent Cold, Dry Cough, and *Snoring*.

2.3. Data Preparation

This stage is an important phase that is often the most time-consuming in machine learning model development, as it involves various technical activities to ensure that the data is clean, consistent, and ready for use in the modelling process [30]. This stage is iterative, allowing for readjustments if obstacles are encountered during the modelling or deployment phases [31]. In this study, the data preparation process includes feature selection by removing irrelevant attributes. Additionally, data cleaning involves removing duplicates and validating the quality of entries, followed by data standardization, which involves converting target labels (*Low, Medium, High*) into a numerical form. Normalization was not applied because the algorithm used is tree-based, which is insensitive to differences in feature scales. Furthermore, categorical numerical features were scaled to provide precise meaning to each feature value. This step aims to enable non-technical users to understand the context of input values when viewing model interpretations using the XAI method.

2.4. Modelling

This stage aims to develop a machine-learning model that can identify patterns in data and generate accurate predictions [30]. This process includes algorithm selection, model training, initial evaluation, and hyperparameter optimization [31]. The machine learning algorithms selected in this study are XGBoost, Random Forest, and LightGBM. The algorithms were selected based on previous studies that demonstrated good accuracy [22]-[24]. All three algorithms belong to the decision tree-based ensemble learning (tree ensemble) [32], [33]. To test the model's generalization to unseen data, validation was performed using the stratified k-fold cross-validation technique for each model. This technique was chosen because it maintains the proportion of target classes in each fold, reduces the risk of overfitting, and has been widely used in various clinical classification studies [27], [34]. Grid Search was applied to the three models —XGBoost, Random Forest, and LightGBM —to evaluate the combination of hyperparameters that yielded the best performance for each model.

2.5. Evaluation

This stage aims to evaluate the model's final performance against test data that was not used during the training process. The results of the system testing are then analyzed to assess the accuracy and effectiveness of the algorithm used [35]. The evaluation was conducted using general metrics in classification, namely accuracy, precision, recall, and F1-score, to assess the model's ability to distinguish between lung cancer risk categories (*Low, Medium, High*). The metrics are calculated based on the values obtained from the confusion matrix, which includes the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) [32]. True Positive (TP) refers to the situation where the model correctly predicts the positive class of an image [36]. When the model incorrectly predicts the positive class of an image, it produces a False Positive (FP) [36]. True Negative (TN) corresponds to cases where the model correctly predicts the negative class of an image [36]. When the model incorrectly predicts the negative class, the result is False Negative (FN) [36]. To obtain these values, see table 1.

Table 1. Confusion Matrix

	Predicted Positive	Predicted Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

Vol. 6, No. 4, Agustus 2025, Page. 2187-2206
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4925

Accuracy is the proportion of total predictions that match the actual values [37]. This value is usually presented as a percentage (%) and represents how well the model can make correct predictions overall. Precision indicates the model's accuracy in classifying data as positive, calculated as the ratio of the number of actual positives correctly predicted to the total number of positives predicted [37]. Meanwhile, recall measures the model's ability to find all actual positive data, thus showing how many positive cases were successfully identified compared to the total number of actual positive cases [37]. The F1-score is the harmonic mean of precision and recall, with values ranging from 0 to 1, where a value of 1 indicates the model's best performance in balancing accuracy and completeness [37]. From table 1, we can derive the formulas for Accuracy, Precision, Recall, and F-1 Score as follows.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1 - score = 2(\frac{(Precision)(Recall)}{Precision + Recall}) \quad (4)$$

2.6. Deployment

This stage aims to implement the prediction model into a system that end-users can utilize. In this study, the implementation was carried out through a web-based *Streamlit* interface that displays lung cancer risk prediction results along with interpretive visualizations. The use of tree ensemble algorithms in this study requires a post-hoc explainability approach to improve understanding of the prediction process [33]. Therefore, the SHAP and LIME methods, which are categorized under feature relevance and visualization techniques, were selected for this analysis. Although not yet applied clinically, this interface is designed to be accessible to non-technical medical personnel, providing clear input options and understandable model explanations. This phase also considers aspects of system sustainability, including ease of model updating, input-output documentation, and potential future integration with clinical systems.

3. RESULT

3.1. Data Understanding

From the dataset there is one feature, Patient Id, as a unique identifier assigned to each patient in the dataset. Table 2 presents descriptive statistics from 1000 observations covering 23 features consisting of risk factors and symptoms related to lung cancer. The feature *Age* shows an average of 37.17 years with a range of 14 to 73 years, indicating that most respondents are of productive age. Several risk features such as *Air Pollution* (3.84), *Alcohol Use* (4.56), *Dust Allergy* (5.16), and *Occupational Hazards* (4.84) have relatively high average values, indicating a relatively significant level of exposure to environmental and lifestyle factors. A similar pattern is observed in the features *of Smoking* and *Passive Smoker*, which have averaged close to 4, reflecting the prevalence of both active and passive smoking behaviors in the sample.

Meanwhile, clinical symptoms such as *Chest Pain* (4.44), *Coughing of Blood* (4.86), and *Shortness of Breath* (4.24) ranked high with relatively high averages, indicating the presence of prominent physical complaints. Conversely, features such as *Snoring* and *Frequent Cold* had lower average values. This distribution pattern suggests that most features are at moderate to high levels, which can generally contribute significantly to the lung cancer risk classification modeling process in the next stage.

Vol. 6, No. 4, Agustus 2025, Page. 2187-2206
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4925

After exploring the dataset, the final process is to identify potential problems in the data, such as missing values, redundancy, and outliers. Based on the results of the examination, we did not find any significant issues related to these three aspects, so the dataset was deemed suitable for further processing in the data preparation stage.

Table 2. Dataset Description

Fitur	Count	Mean	Std	Min	25%	50%	75%	Max
Age	1000	37.17	12.01	14	27.75	36	45	73
Gender	1000	1.40	0.49	1	1	1	2	2
Air Pollution	1000	3.84	2.03	1	2	3	6	8
Alcohol Use	1000	4.56	2.62	1	2	5	7	8
Dust Allergy	1000	5.16	1.98	1	4	6	7	8
Occupational Hazards	1000	4.84	2.11	1	3	5	7	8
Genetic Risk	1000	4.58	2.13	1	2	5	7	7
Chronic Lung Disease	1000	4.38	1.85	1	3	4	6	7
Balanced Diet	1000	4.49	2.14	1	2	4	7	7
Obesity	1000	4.46	2.12	1	3	4	7	7
Smoking	1000	3.95	2.50	1	2	3	7	8
Passive Smoker	1000	4.20	2.31	1	2	4	7	8
Chest Pain	1000	4.44	2.28	1	2	4	7	9
Coughing of Blood	1000	4.86	2.43	1	3	4	7	9
Fatigue	1000	3.86	2.24	1	2	3	5	9
Weight Loss	1000	3.86	2.21	1	2	3	6	8
Shortness of Breath	1000	4.24	2.29	1	2	4	6	9
Wheezing	1000	3.78	2.04	1	2	4	5	8
Swallowing Difficulty	1000	3.75	2.27	1	2	4	5	8
Clubbing of Finger Nails	1000	3.92	2.39	1	2	4	5	9
Frequent Cold	1000	3.54	1.83	1	2	3	5	7
Dry Cough	1000	3.85	2.04	1	2	4	6	7
Snoring	1000	2.93	1.47	1	2	3	4	7

3.2. Data Preparation

Based on the descriptive analysis of each feature, we decided to remove the *Patient Id* feature from the dataset because it was not relevant to the predictive purpose of this study. After this process, we validated the dataset for outliers and redundant data. The examination found that the dataset did not contain any outliers, but there were 848 duplicate entries. Therefore, we cleaned the data, resulting in 152 unique entries.

Feature importance analysis was performed to evaluate the contribution of each feature to the classification ability of the Random Forest model in table 3. The results show that features such as Coughing of Blood (0.113), Wheezing (0.075), and Dust Allergy (0.072) have the highest weight in influencing the prediction output. Conversely, features such as Gender and Age showed very low contributions, with values below 0.01.

To determine the optimal number of features used in modeling, a cumulative feature importance calculation was performed. The features were sorted based on their importance value in descending order, then summed up until their accumulated contribution reached a threshold of 90%. From the results shown in figure 2, it was found that the first 17 features already covered a cumulative contribution of ≥90%, which is considered sufficient to maintain accuracy without overly increasing model complexity. This approach is beneficial for simplifying the model, improving computational efficiency, and minimizing the risk of overfitting, especially in datasets with many redundant or non-important features.

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 4, Agustus 2025, Page. 2187-2206
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4925

Table 3. Feature Importance

Feature	Importance
Coughing of Blood	0.113294
Wheezing	0.074536
Dust Allergy	0.072444
Passive Smoker	0.068753
Balanced Diet	0.063387
Obesity	0.063196
Fatigue	0.055607
Alcohol Use	0.045861
Occupational Hazards	0.044680
Air Pollution	0.043680
Chest Pain	0.043592
Shortness of Breath	0.042337
Smoking	0.037167
Frequent Cold	0.037080
Genetic Risk	0.033826
Clubbing of Finger Nails	0.032283
Swallowing Difficulty	0.030634
Weight Loss	0.029044
Snoring	0.027186
Dry Cough	0.019841
Chronic Lung Disease	0.015239
Age	0.006108
Gender	0.000226

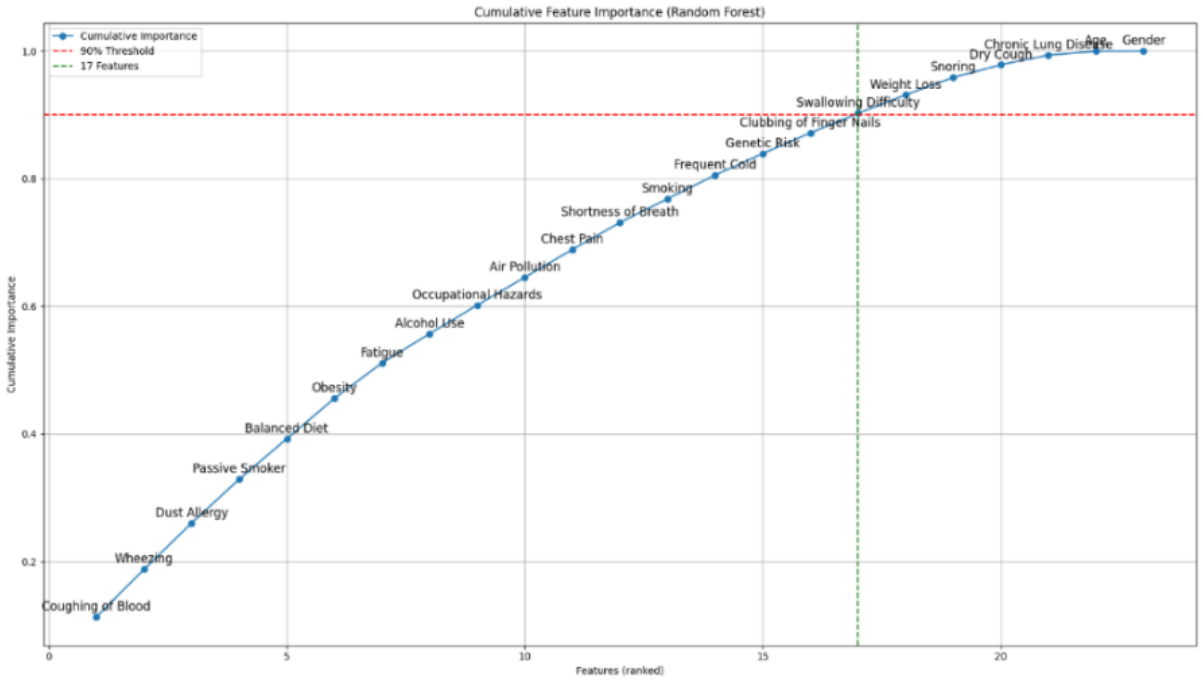


Figure 2. Cumulative Feature Importance

Vol. 6, No. 4, Agustus 2025, Page. 2187-2206
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4925

Furthermore, table 2 shows inconsistencies in the naming of several features in the dataset. To improve readability and consistency, we normalized the feature names by adjusting the use of capital letters. Three features were adjusted—namely, *Alcohol use* to *Alcohol Use* and *OccuPational Hazards* to *Occupational Hazards*.

In the same stage, the feature *Level* is included in the object categorical with three categorical values, namely *Low*, *Medium*, and *High*. To facilitate processing by machine learning algorithms that require numeric input, the three categories were labeled. The results of this process show that the label *Low* is coded as "0", *Medium* as "1", and *High* as "2".

Similar adjustments to numerical categorical features were made based on a literature review, as shown in table 4 for risk factors, table 5.1 and table 5.2 for symptoms. The existing literature review has explained and proven that these features are influential in lung cancer.

Table 4. Identification Categorical Numeric Risk Factors

Feature	Scale	Category
Air Pollution [38], [39]	1-8	0–5 μg/m³; 6–10 μg/m³; 11–15 μg/m³; 16–20 μg/m³; 21–25 μg/m³; 26–30 μg/m³; 31–40 μg/m³; >41 μg/m³
Alcohol Consumption [40]-[42]	1-8	0 g/day (0 drinks/day); 1–5 g/day (0.5 drinks/day); 6–12 g/day (1 drink/day); 13–24 g/day (1–2 drinks/day); 25–35 g/day (2–3 drinks/day); 36–50 g/day (3–4 drinks/day); 51–75 g/day (4–6 drinks/day); >75 g/day (>6 drinks/day)
Dust Allergy [43]-[45]	1-8	No dust allergy or exposure; Mild, occasional dust contact; Moderate indoor allergy, controlled; Regular allergy + mild symptoms; Diagnosed + moderate asthma; Severe allergy + poor control; High occupational dust exposure; Chronic high exposure + no control
Occupational Hazards [46]-[51]	1-8	No occupational exposure; Occasional low exposure; Regular low exposure (e.g. drivers); Moderate exposure to one carcinogen; Moderate-high exposure (e.g. diesel); High exposure (asbestos, etc.); High-risk job >10 years; Multiple exposures + smoking
Genetic Risk [52], [53]	1-7	No family history of lung cancer; Distant relative (2nd-degree); 1st-degree relative >60 y.o.; 1st-degree relative <60 y.o.; Two 1st-degree relatives; Two early-onset cases; Multiple relatives/genetic mutation
Balanced Diet [54]-[56]	1-7	Optimal: high plant, low red meat; (Med/PDI); Very high LLDS, low processed; Plant-rich, low-fat; High whole grains, non-oily fish; Moderate fruit/veg + some meat; Occasional fruit/veg; Poor (high red meat, low veg)
Obesity [57]	1-7	Very underweight; Underweight; Normal; Overweight; Obesity I; Obesity II; Obesity III
Smoking [58], [59]	1-8	Never smoked / <100 lifetime; Very light exposure; Light smoker; Moderate smoker; Heavy smoker; Very heavy; Extremely heavy; Former heavy smoker >15 years
Passive Smoker [60], [61]	1-8	No exposure; Occasional exposure; Household <10 yrs; Household >10 yrs; Workplace <20 yrs; Workplace ≥20 yrs or 2 sources; Childhood + adult + work; Chronic, multi-source >20 yrs

Vol. 6, No. 4, Agustus 2025, Page. 2187-2206
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4925

Table 5.1. Identification Categorical Numeric Symptoms

Feature	Scale	Category
Chest Pain [62], [63]	1-9	No chest pain; Very mild occasional ache (e.g., on deep breath/exertion); Mild discomfort with moderate activity; Mild-moderate pain during daily tasks; Moderate pain, impacting some activities; Moderate-severe daily pain; Severe pain (limits activity, MDASI ≥7); Very severe pain nearly daily; Debilitating pain at rest (MDASI 9–10)
Coughing of Blood [63]	1-9	No coughing up blood; Very minimal: streaks only once; Mild traces (<1 tsp, once/month); Occasional small streaks (~1 tsp, a few times/month); Frequent mild (daily streaks, <1 tsp); Moderate daily bleeding (1–2 tsp); Daily significant bleeding (2–3 tsp); Heavy bleeding (>3 tsp), distressing; Massive hemoptysis (emergency-level)
Fatigue [63], [64]	1-9	No fatigue; Very mild fatigue during unusual exertion; Mild fatigue after activity; Mild-moderate fatigue—frequent but manageable; Moderate fatigue—daily, noticeable; Moderate-severe fatigue—daily, affects chores; Severe fatigue (MDASI ≥7)—limits most activities; Very severe—daily exhaustion, rest needed; Debilitating fatigue (MDASI 9–10)—unable to function
Shortness of Breath [63], [65], [66]	1-9	No breathlessness; Very mild on strenuous activity; Mild on moderate exertion; Shortness during daily tasks; Moderate SOB with normal activities; Daily, persistent SOB; Severe SOB during moderate tasks (MDASI ≥7); Very severe SOB, at rest; Debilitating SOB with minimal effort
Wheezing [63]	1-8	No wheezing; Rare, during colds or exercise; Wheezing <1x/week; Wheezing several times/week; Frequent, without infection; Daily wheezing during activities; Severe wheezing at rest; Distressing, continuous wheezing

Table 5.2. Identification Categorical Numeric Symptoms cont.

Feature	Scale	Category
Swallowing Difficulty [63], [67], [68]	1-8	No swallowing difficulty; Occasional throat discomfort with solids; Difficulty with solids once or twice/week; Needs semi-solid/liquid diet occasionally; Semi-solid/liquid diet daily; Difficulty swallowing liquids; Can only swallow saliva, frequent choking; Unable to swallow solids/liquids (needs NGT)
Clubbing of Finger Nails [63], [69]	1-9	No clubbing; normal Lovibond angle; Slightly curved nail tips; Mild rounding with partial Schamroth loss; Persistent rounding; angle >180°; Obvious bulbous enlargement; Clubbing + joint pain/swelling; Clubbing + periostosis on imaging; Extreme clubbing + systemic signs; Severe clubbing + cancer-related features
Frequent Cold [63], [70]	1-7	No respiratory infections in past year; 1 mild cold/year; 2–3 mild infections/year; 4–5 infections/year; ≥6 infections or ≥1 needing antibiotics; ≥2 moderate infections/year; 3+ hospital-treated infections

3.3. Modelling

After the data preparation process is complete and the dataset is ready for analysis, the next step is to apply machine learning algorithms. The process of modelling and also the evaluation as seen in figure

Vol. 6, No. 4, Agustus 2025, Page. 2187-2206
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4925

3. In this study, three algorithms were selected to build predictive models, namely XGBoost, RF, and LightGBM. Before that, to ensure the model runs optimally, hyperparameter tuning was performed using the Grid Search Cross Validation method. This method searches for the best parameter combination based on model performance on the validation data. Table 6 summarizes the best hyperparameter configurations for each algorithm based on the tuning results.

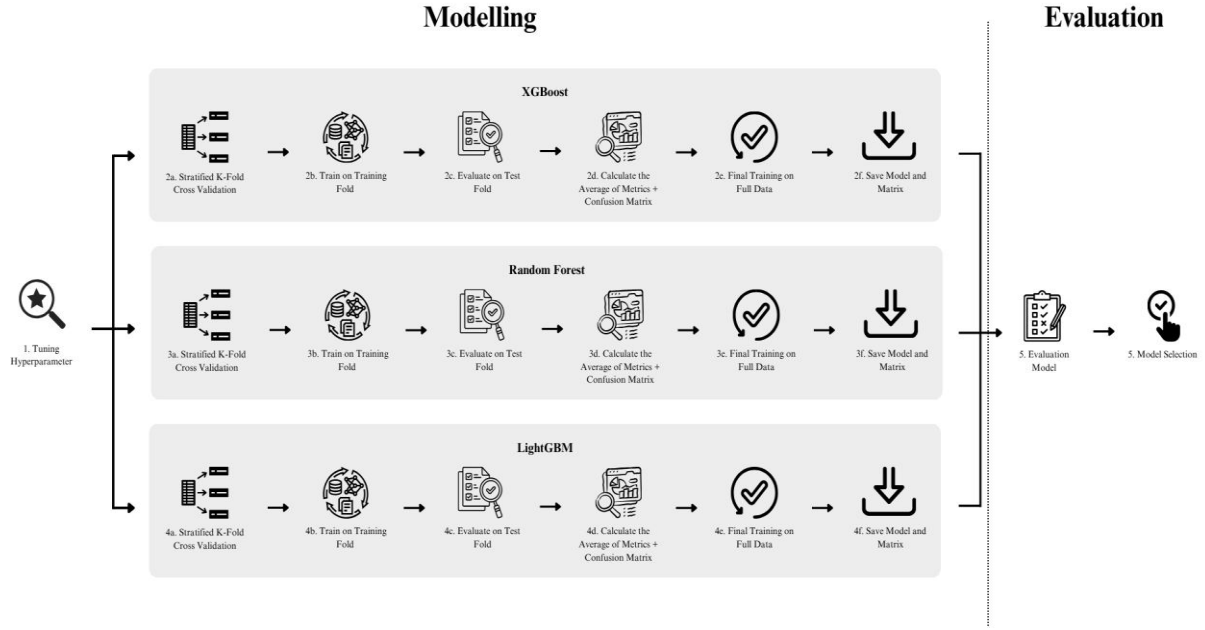


Figure 3. Modelling and Evaluation Process

Table 6. Hyperparameter Tunning

XGBoost [15]	*'colsample_bytree': 0.8 'gamma': 0 'learning_rate': 0.1 'max_depth': 6 'min_child_weight': 1 'n_estimators': 200 'scale_pos_weight': 1 'subsample': 0.8*	**RF** [15]	*bootstrap': True 'max_depth': None, 'max_features': 'sqrt' 'min_samples_leaf': 2 'min_samples_split': 2 'n_estimators': 100*	**LightGBM** [15]	*'colsample_bytree': 0.8 'learning_rate': 0.1 'max_depth': 6 'min_child_weight': 1 'n_estimators': 100 'num_leaves': 31 'subsample': 0.8*

To obtain optimal results and avoid the risk of overfitting, a Stratified K-Fold Cross Validation approach with a value of k equal to 5 was used. This approach was chosen to ensure that the data distribution in each fold remained balanced between target classes. In each iteration, the model was trained using four parts of the data and validated on the remaining part until all data had been used as validation data.

3.4. Evaluation

Table 7. The Average of Algorithm Results

	Random Forest	XGBoost	LightGBM
Accuracy (%)	**97.38**	96.71	96.71
Precision (%)	**97.73**	97.22	97.17
Recall (%)	**97.26**	96.59	96.52
F1-Score (%)	**97.30**	96.69	96.57

Vol. 6, No. 4, Agustus 2025, Page. 2187-2206
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4925

To validate model stability across the dataset, we also performed Stratified 5-Fold Cross Validation. Although table 7 presents the average scores, each fold's metrics were consistently high. This consistency indicates that the model is robust and generalizes well to unseen data. The evaluation results, as seen in table 7, show that Random Forest produced the best performance with an accuracy of 97.38%, precision of 97.73%, recall of 97.26%, and F1-score of 97.30%. These figures are higher than those of XGBoost and LightGBM, which achieved an accuracy of 96.71%. In addition, from figure 4 the average confusion matrix shows that Random Forest can classify all classes with a very low error rate and balanced prediction distribution. Based on these results, the Random Forest model was selected as the main model in this study. This selection was made due to its consistently superior metric performance and stability in processing data across folds. As a result, Random Forest is considered the most suitable model for use in the lung cancer risk classification system in this study. In addition, the average confusion matrix shows that Random Forest can classify all classes with a very low error rate and balanced prediction distribution.

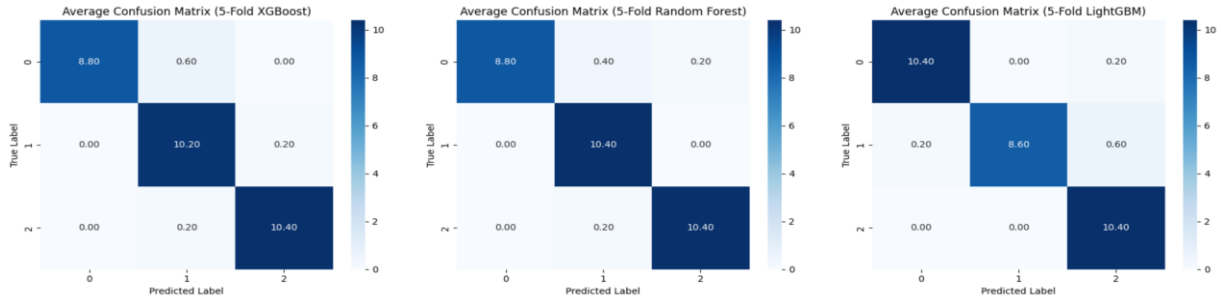


Figure 4. The Average of Matrix Confusion

Based on these results, the Random Forest model was selected as the main model in this study. This selection was made due to its consistently superior metric performance and stability in processing data across folds. As a result, Random Forest is considered the most suitable model for use in the lung cancer risk classification system in this study.

3.5. Deployment

In this study, the Random Forest classification model that has undergone training and validation processes was applied to a graphical user interface (GUI)-based application using the Streamlit framework, as shown in figure 5. This system allows patient data input based on risk factors and symptoms. Based on the input data, the model predicts whether the patient falls into the *Low, Medium,* or *High* category. The prediction interpretation is provided through two Explainable AI (XAI) approaches, namely SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations), which are then reinforced with automatic narrative explanations from the Chat-GPT model [26].

Figure 6-10 presents the risk prediction result for an 8-year-old girl presenting with severe hemoptysis (±110 cc/day), nonproductive cough, shortness of breath, and wheezing in the left hemithorax that had been ongoing for three months [71]. The patient also experienced significant weight loss (5 kg in 2 weeks) without a history of smoking, exposure to pollution, or genetic predisposition. Radiology revealed a mediastinal mass with compression of the left bronchus and air trapping. Bronchoscopy revealed total obstruction of the left bronchus, and biopsy confirmed stage T1N0M0 lung adenocarcinoma. No metastasis was detected, and lobectomy was performed without chemotherapy. The patient is currently under follow-up without recurrence.

From the result figure 6 presents the model prediction output, which classifies the patient as Low risk with a probability of 84.89%. To provide transparency, SHAP values are visualized in figure 7, indicating that Wheezing (2), Alcohol Use (1), and Obesity (3) contribute the most to the risk score. The SHAP explanation in figure 8 reinforces these findings by detailing the influence of each feature, such as

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 4, Agustus 2025, Page. 2187-2206
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4925

mild wheezing having the highest SHAP value (+0.11), aligning with known clinical evidence. Figure 9 shows the LIME explanation, in which the most influential features for ruling out the Medium-risk class are Wheezing (2.00), Dust Allergy (1.00), and Alcohol Use (1.00). Figure 10 provides a narrative summary confirming the alignment between SHAP and LIME, increasing the interpretability and clinical trust in the model's prediction.

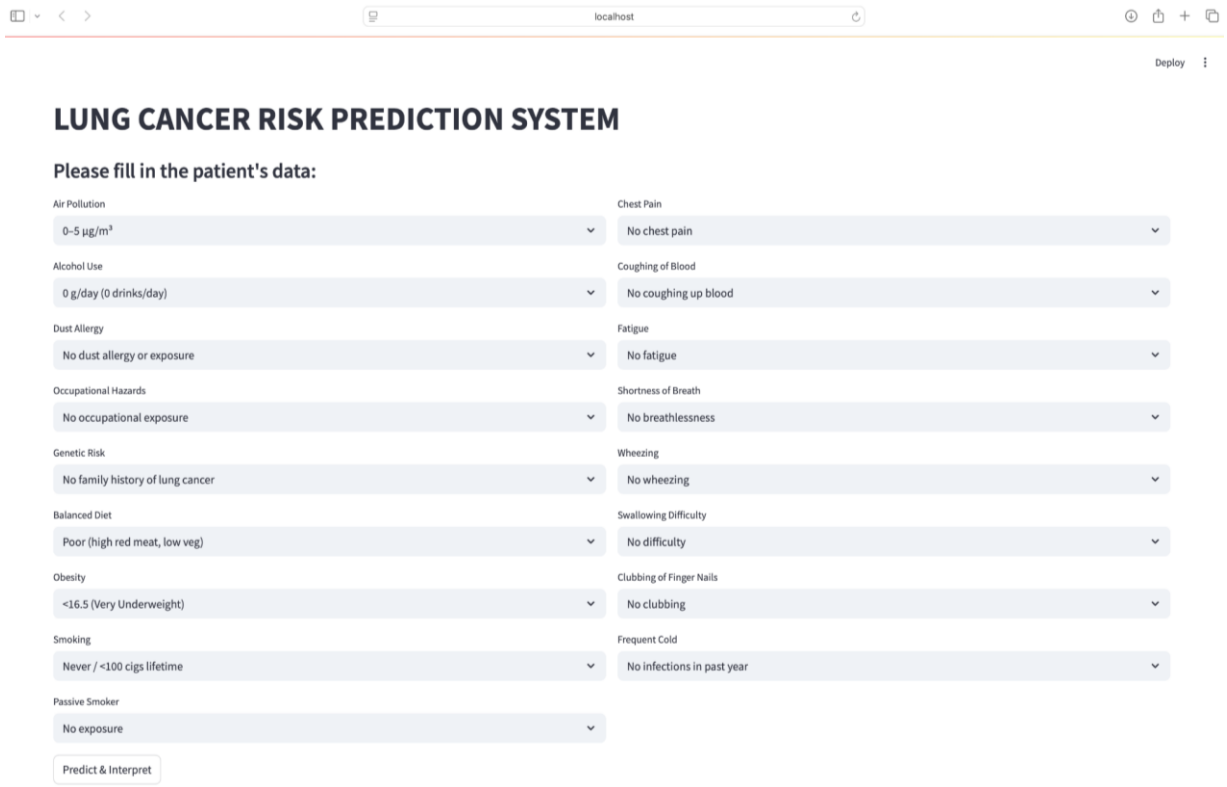


Figure 5. Graphic User Interface

Figure 6. Prediction Probability of System

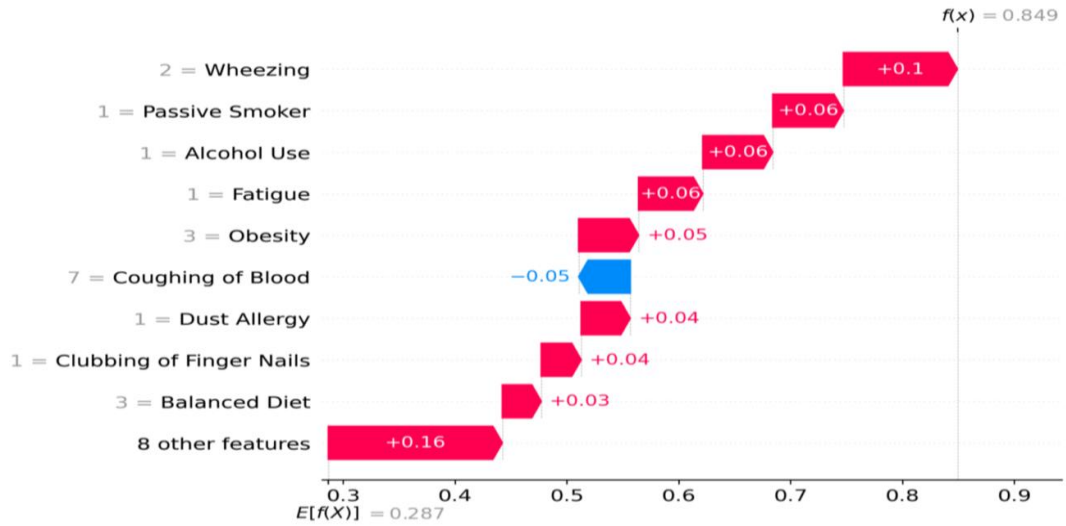


Figure 7. SHAP Waterfall Interpretation

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 4, Agustus 2025, Page. 2187-2206
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4925

SHAP Explanation

Main Supporting Features

- **Wheezing (2)** – This symptom shows the strongest positive SHAP value (+0.10), suggesting that even mild wheezing contributes significantly to lung cancer risk prediction.
- **Passive Smoker (1)** – Despite being minimal exposure, this factor still contributes positively (+0.06), aligning with evidence of secondhand smoke as a cancer risk.
- **Alcohol Use (1)**, **Fatigue (1)**, and **Obesity (3)** – Each adds moderate positive SHAP values (+0.05–0.06), reflecting that lifestyle and general health indicators slightly raise predicted risk.

Opossing Features

- **Coughing of Blood (7)** – Surprisingly contributes negatively (–0.05), possibly due to its isolated occurrence among otherwise low-risk factors.
- **Dust Allergy (1)** and **Balanced Diet (3)** – Minor protective effects (each +0.03–0.04), consistent with lower inflammatory or dietary burden.

Clinical Summary

Despite the presence of one critical symptom (*hemoptysis*), the overall low-risk profile—including no smoking history, minimal passive exposure, and mild symptoms—resulted in a confident **Low** risk classification (84.89%). The model suggests that isolated red flags may not outweigh a generally favorable risk background.

Figure 8. Chat-GPT Explanation of SHAP Interpretation

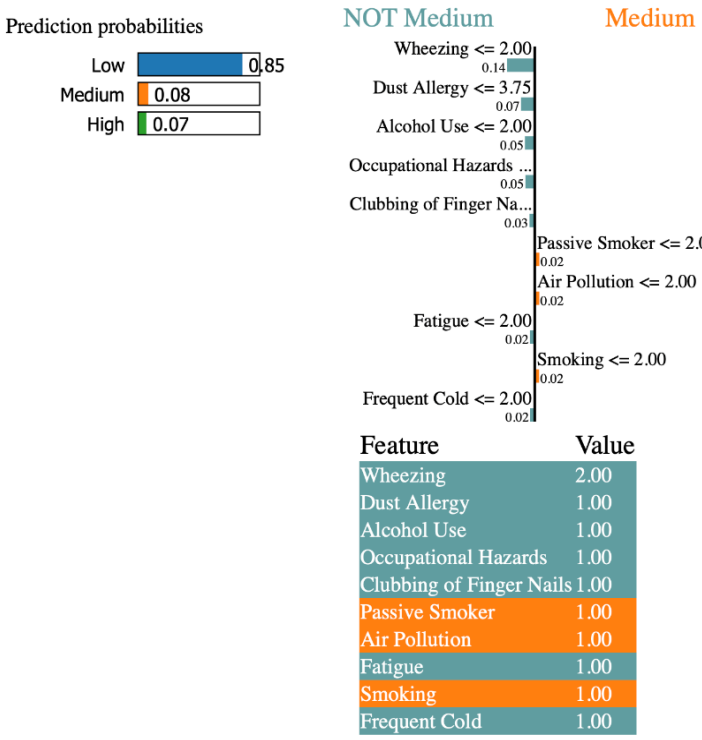
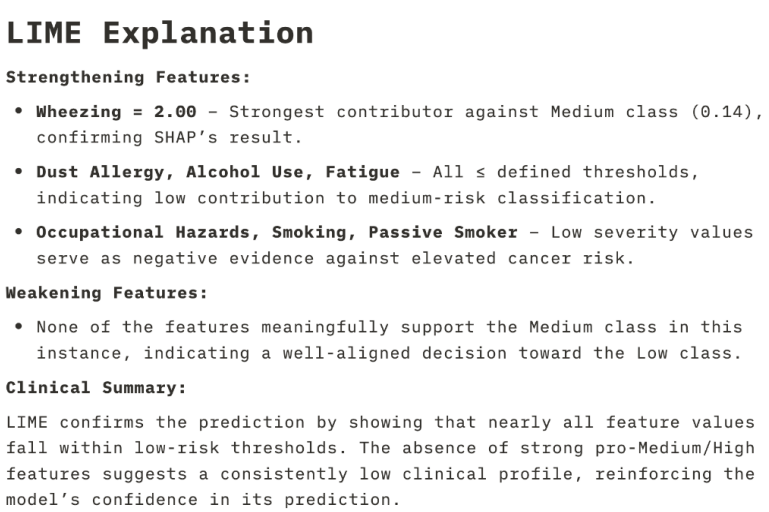


Figure 9. LIME Interpretation

Vol. 6, No. 4, Agustus 2025, Page. 2187-2206
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4925



LIME Explanation

Strengthening Features:

- **Wheezing = 2.00** – Strongest contributor against Medium class (0.14), confirming SHAP's result.
- **Dust Allergy, Alcohol Use, Fatigue** – All ≤ defined thresholds, indicating low contribution to medium-risk classification.
- **Occupational Hazards, Smoking, Passive Smoker** – Low severity values serve as negative evidence against elevated cancer risk.

Weakening Features:

- None of the features meaningfully support the Medium class in this instance, indicating a well-aligned decision toward the Low class.

Clinical Summary:

LIME confirms the prediction by showing that nearly all feature values fall within low-risk thresholds. The absence of strong pro-Medium/High features suggests a consistently low clinical profile, reinforcing the model's confidence in its prediction.

Figure 10. Chat-GPT Explanation of LIME Interpretation

4. DISCUSSION

Previous studies have explored the effectiveness of machine learning algorithms in predicting lung cancer risk. A 2022 study demonstrated that ensemble methods such as Random Forest and Rotation achieved 97.1% for the accuracy in the binary classification of lung cancer [22]. These findings highlight the accuracy of tree-based models in recognizing complex clinical patterns. Although this study was limited to two classes (positive and negative cancer), the approach remains relevant in this study, which develops a multi-class classification based on risk levels.

A 2024 study evaluated several algorithms, including XGBoost and LightGBM, with XGBoost providing the best performance (97.50% accuracy) in lung cancer prediction based on symptom and lifestyle data [24]. However, their study did not emphasize the interpretability aspect of the model. This study complements this shortcoming by not only comparing performance but also integrating Explainable AI (XAI) techniques to support model decision transparency.

In the same year, 2024, several algorithms were explored, including SVM, RF, DT, and KNN. In this study, the importance of interpretability in multi-class lung cancer risk classification was emphasized [25]. They also used LIME to explain decisions made by Random Forest and SVM models, concluding that local-global visualizations can enhance trust among non-technical users. This research expands this approach by adding automatic explanations based on natural language through a GPT model, enabling access for users unfamiliar with technical visualizations.

Furthermore, In 2020 highlight that an ideal Explainable AI system should be able to provide interpretations that are understandable, trustworthy, and actionable [33]. The approach applied in this study aligns with these principles through SHAP and LIME-based visualizations, complemented by automatic explanations that both medical professionals and patients can understand.

Based on this conceptual framework, this study evaluates three machine learning algorithms—Random Forest, XGBoost, and LightGBM—for lung cancer risk classification. The evaluation was validated with 5-fold cross validation. The results show that Random Forest achieves the highest performance, with an accuracy of 97.38%, precision of 97.73%, recall of 97.26%, and an F1-score of 97.30%.

This model not only excels in terms of metric performance but also demonstrates stability in classification distribution on the confusion matrix, making it the primary model proposed. To transparently explain prediction results, two visual interpretation methods were applied: SHAP and LIME. System evaluation was conducted on the medical report of an 8-year-old child patient. SHAP visualization

Vol. 6, No. 4, Agustus 2025, Page. 2187-2206
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4925

shows that features such *Wheezing*, *Passive Smoker*, and *Alcohol Use* contribute positively to the low-risk classification while *Coughing of Blood* rejects the prediction. LIME visualization on a case of a girl with severe symptoms but no classic risk exposure shows that other high-risk features are not dominant, supporting the *Low* class prediction.

As a complement, the system includes GPT-based explanations that generate automatic narratives from the SHAP and LIME results. These narratives address three main points: supporting features, opposing features, and brief clinical interpretations. The combination of visual and narrative elements, the prediction system is not only interpretable by researchers but also understandable by medical staff without deep technical backgrounds. This multimodal approach reinforces the principles of transparency and accountability in the use of clinical AI.

Although the developed system shows promising performance and interpretability, several limitations need to be considered. First, the dataset used is sourced from open-source and non-clinical sources, so it may not fully represent the diversity of patient data in the real world. Second, although the system has been tested on the medical report of an 8-year-old child with severe respiratory symptoms and a diagnosis of lung adenocarcinoma, this trial is still illustrative and cannot replace actual clinical validation. Third, the user interface and automatic narratives generated by the GPT model, although intended to facilitate understanding for non-technical users, have not been directly evaluated by healthcare professionals to assess their clarity, accuracy, or usefulness in a clinical context. Therefore, further research is needed to test this model on real patient data, involving evaluation by medical professionals, and exploring its potential integration into decision-making workflows in healthcare services.

5. CONCLUSION

This study demonstrates the potential of integrating machine learning and explainable artificial intelligence (XAI) to support early detection and risk classification of lung cancer. The results show that Random Forest achieves the highest performance, with an accuracy of 97.38%, precision of 97.73%, recall of 97.26%, and an F1-score of 97.30%. By leveraging Random Forest as the primary classifier, the system achieved high predictive performance while maintaining stability across different data partitions. SHAP and LIME are the methods deployed to provide transparent and accessible explanations of the model's predictions for making the outputs more understandable for non-technical. The integration of the GPT-based model allows the system to communicate critical insights in a human-readable format. This combination reinforces trust in AI-driven decisions. Then, the prototype of GUI can demonstrate a practical, interpretable, and scalable approach for risk stratification in lung cancer, aligning with the broader goals of responsible and explainable medical AI deployment.

Further research is recommended to test the system's application on actual clinical data, involving assessments by healthcare professionals, and evaluating the level of usability and effectiveness of the interface in real-world screening contexts. Such efforts are important to ensure that this AI-based system can be effectively and reliably applied to support public health policies and national screening programs.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest between the authors or with research objects in this paper.

ACKNOWLEDGEMENT

The authors express their sincere gratitude to all parties who have provided valuable support and contributions, enabling the successful completion of this journal.

Vol. 6, No. 4, Agustus 2025, Page. 2187-2206
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4925

REFERENCES

[1] A. Mahmood and R. Srivastava, "CHAPTER 3 - Etiology of cancer," in *Understanding Cancer*, B. Jain and S. Pandey, Eds. Academic Press, 2022, pp. 37–62, doi: 10.1016/B978-0-323-99883-3.00008-1.

[2] F. Bray, M. Laversanne, H. Sung, J. Ferlay, R. L. Siegel, I. Soerjomataram, and A. Jemal, "Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA Cancer J. Clin.*, vol. 74, no. 4, pp. 362–387, Apr. 2024, doi: 10.3322/caac.21834.

[3] K. C. Thandra, A. Barsouk, K. Saginala, J. S. Aluru, and A. Barsouk, "Epidemiology of lung cancer," *Contemp. Oncol. (Pozn)*, vol. 25, no. 1, pp. 45–52, 2021, doi: 10.5114/wo.2021.103829.

[4] O. D. Asmara *et al*., "Lung cancer in Indonesia," *J. Thorac. Oncol.*, vol. 18, no. 9, pp. 1134–1145, 2023, doi: 10.1016/j.jtho.2023.06.010.

[5] Kementerian Kesehatan Republik Indonesia, *Rencana Kanker Nasional 2024–2034*, Jakarta: Direktorat P2PTM, 2024. [Online]. Available: https://www.iccp-portal.org/sites/default/files/plans/Rencana_Kanker_Nasional_2024-2034.pdf

[6] A. S. Ahmad and A. M. Mayya, "A new tool to predict lung cancer based on risk factors," *Heliyon*, vol. 6, no. 2, p. e03402, 2020, doi: 10.1016/j.heliyon.2020.e03402.

[7] S. M. Sakthisankaran, D. Sakthipriya, and M. Swamivelmanickam, "Health risks associated with tobacco consumption in humans: An overview," *J. Drug Deliv. Ther.*, vol. 14, no. 5, 2024.

[8] A. Agustí *et al*., "Global Initiative for Chronic Obstructive Lung Disease 2023 Report: GOLD Executive Summary," *Eur. Respir. J.*, vol. 61, no. 4, 2300239, 2023, doi: 10.1183/13993003.00239-2023.

[9] L. S. Flor, J. A. Anderson, N. Ahmad, *et al*., "Health effects associated with exposure to secondhand smoke: a Burden of Proof study," *Nat. Med.*, vol. 30, pp. 149–167, 2024, doi: 10.1038/s41591-023-02743-4.

[10] O. I. Onwurah, "A data analysis of the correlation between smoking and lung cancer," *SSRN*, May 11, 2025, doi: 10.2139/ssrn.5250309.

[11] Y. Huang *et al*., "Air pollution, genetic factors, and the risk of lung cancer: A prospective study in the UK Biobank," *Am. J. Respir. Crit. Care Med.*, vol. 204, no. 7, pp. 817–825, 2021, doi: 10.1164/rccm.202011-4063OC.

[12] M. I. D. Rakasiwi, W. Prasetya, I. Riyatno, *et al*., "Starting early palliative care for suspected lung cancer patient: A case series from resource-limited setting in Indonesia," *Rwanda Med. J.*, vol. 80, no. 4, pp. 5–9, 2023, doi: 10.4314/rmj.v81i2.1.

[13] S. Yulianti, M. A. Budiman, and M. Amri, "Utilization of radiological techniques in early diagnosis of lung cancer," *Int. J. Eng. Emerg. Technol. (IJEET)*, vol. 2, no. 2, pp. 191–197, Mar. 2024, doi: 10.61991/ijeet.v2i2.35.

[14] U. Chandran, J. Reps, R. Yang, A. Vachani, F. Maldonado, and I. Kalsekar, "Machine learning and real-world data to predict lung cancer risk in routine care," *Cancer Epidemiol. Biomarkers Prev.*, vol. 32, no. 3, pp. 337–343, 2023, doi: 10.1158/1055-9965.EPI-22-0873.

[15] L. Swinckels *et al*., "The use of deep learning and machine learning on longitudinal electronic health records for the early detection and prevention of diseases: Scoping review," *J. Med. Internet Res.*, vol. 26, e48320, 2024, doi: 10.2196/48320.

[16] A. Houston, S. Williams, W. Ricketts, *et al*., "Automated derivation of diagnostic criteria for lung cancer using natural language processing on electronic health records: A pilot study," *BMC Med. Inform. Decis. Mak.*, vol. 24, p. 371, 2024, doi: 10.1186/s12911-024-02790-y.

[17] A. N. Fatyandri, B. Guo, and E. S. Muchsinati, "Impact of artificial intelligence and human resource management on leadership organization performance," *J. Tek. Manaj. Inform.*, vol. 10, no. 2, pp. 123–132, 2024, doi: 10.26905/jtmi.v10i2.14060.

[18] N. Kühl, M. Schemmer, M. Goutier, *et al*., "Artificial intelligence and machine learning," *Electron. Markets*, vol. 32, pp. 2235–2244, 2022, doi: 10.1007/s12525-022-00598-0.

[19] S. Gondhowiardjo *et al*., "Five-year cancer epidemiology at the national referral hospital: Hospital-based cancer registry data in Indonesia," *JCO Glob. Oncol.*, vol. 7, pp. 190–203, 2021, doi: 10.1200/GO.20.00155.

Vol. 6, No. 4, Agustus 2025, Page. 2187-2206
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4925

[20] V. Hassija, V. Chamola, A. Mahapatra, *et al*., "Interpreting black-box models: A review on explainable artificial intelligence," *Cogn. Comput.*, vol. 16, pp. 45–74, 2024, doi: 10.1007/s12559-023-10179-8.

[21] N. R. Nuraeda, M. Liebenlito, and T. E. Sutanto, "Explainable sentiment analysis pada ulasan aplikasi Shopee menggunakan Local Interpretable Model-agnostic Explanations," *Indones. J. Comput. Sci.*, vol. 13, no. 3, 2024, doi: 10.33022/ijcs.v13i3.3870.

[22] E. Dritsas and M. Trigka, "Lung cancer risk prediction with machine learning models," *Big Data Cogn. Comput.*, vol. 6, no. 4, p. 139, 2022, doi: 10.3390/bdcc6040139.

[23] M. Mamun, A. Farjana, M. A. Mamun, and M. S. Ahammed, "Lung cancer prediction model using ensemble learning techniques and a systematic review analysis," in *2022 IEEE World AI IoT Congress (AIIoT)*, 2022, pp. 187–193, doi: 10.1109/AIIOT54504.2022.9817326.

[24] M. M. R. Sweet, M. P. Ahmed, M. A. S. Mozumder, *et al*., "Comparative analysis of machine learning techniques for accurate lung cancer prediction," *Am. J. Eng. Technol.*, vol. 6, no. 9, pp. 92–103, Sep. 2024, doi: 10.37547/tajet/Volume06Issue09-11.

[25] R. K. Pathan, I. J. Shorna, M. S. Hossain, *et al*., "The efficacy of machine learning models in lung cancer risk prediction with explainability," *PLoS ONE*, vol. 19, no. 6, p. e0305035, 2024, doi: 10.1371/journal.pone.0305035.

[26] L. Dwiyanti, N. Nambo, and N. Hamid, "Leveraging Explainable Artificial Intelligence (XAI) for expert interpretability in predicting rapid kidney enlargement risks in autosomal dominant polycystic kidney disease (ADPKD)," *AI*, vol. 5, no. 4, pp. 2037–2065, 2024, doi: 10.3390/ai5040100.

[27] V. Singh, A. Singh, and K. Joshi, "Fair CRISP-DM: Embedding fairness in machine learning (ML) development life cycle," in *Proc. 55th Hawaii Int. Conf. on System Sciences (HICSS)*, 2022, pp. 260–269. [Online]. Available: https://aisel.aisnet.org/hicss-55/da/algorithmic_fairness/3.

[28] M. Cazacu and E. Titan, "Peculiarities of providing psychological assistance to abused children," *BRAIN: Broad Research in Artificial Intelligence and Neuroscience*, vol. 11, no. 2Sup1, pp. 99–106, 2020, doi: 10.18662/brain/11.2Sup1/97.

[29] TheDevastator, "Lung Cancer Prediction," *Kaggle Datasets*, 2020. [Online]. Available: https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link.

[30] M. Elkabalawy, A. Al-Sakkaf, E. M. Abdelkader, and G. Alfalah, "CRISP-DM-based data-driven approach for building energy prediction utilizing indoor and environmental factors," *Sustainability*, vol. 16, no. 17, p. 7249, 2024, doi: 10.3390/su16177249.

[31] S. Studer, T. B. Bui, C. Drescher, A. Hanuschkin, L. Winkler, S. Peters, and K.-R. Müller, "Towards CRISP-ML(Q): A machine learning process model with quality assurance methodology," *Mach. Learn. Knowl. Extract.*, vol. 3, no. 2, pp. 392–413, 2021, doi: 10.3390/make3020020.

[32] M. A. Muslim, T. L. Nikmah, D. A. A. Pertiwi, Subhan, Jumanto, Y. Dasril, and Iswanto, "New model combination meta-learner to improve accuracy prediction P2P lending with stacking ensemble learning," *Inform. Sci. Wound. Appl.*, vol. 20, p. 200204, 2023, doi: 10.1016/j.iswa.2023.200204.

[33] A. B. Arrieta *et al.*, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, 2020, doi: 10.1016/j.inffus.2019.12.012.

[34] A. Manzoor, M. A. Qureshi, E. Kidney, and L. Longo, "A review on machine learning methods for customer churn prediction and recommendations for business practitioners," *IEEE Access*, vol. 12, pp. 70434–70463, 2024, doi: 10.1109/ACCESS.2024.3402092.

[35] N. Arifin, C. N. Insani, M. Milasari, J. Rusman, S. Upa, and M. S. A. Utama, "Classification of helmet and vest usage for occupational safety monitoring using backpropagation neural network," *J. Tek. Inform. (JUTIF)*, vol. 6, no. 3, pp. 1255–1266, Jun. 2025, doi: 10.52436/1.jutif.2025.6.3.4781.

[36] K. Ali, Z. A. Shaikh, A. A. Khan, and A. A. Laghari, "Multiclass skin cancer classification using EfficientNets – a first step towards preventing skin cancer," *Neuroscience Informatics*, vol. 2, no. 4, Art. no. 100034, Dec. 2022, doi: 10.1016/j.neuri.2021.100034.

[37] R. D. Marzuq, S. A. Wicaksono, and N. Y. Setiawan, "Prediksi kanker paru-paru menggunakan algoritme Random Forest Decision Tree," *J. Pengemb. Teknol. Inf. Ilmu Komput.*, vol. 7, no. 7,

Vol. 6, No. 4, Agustus 2025, Page. 2187-2206
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4925

pp. 3448–3456, 2023. [Online]. Available: https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/12964.

[38] World Health Organization, "WHO global air quality guidelines: Particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide," Geneva: World Health Organization, 2021. Licence: CC BY-NC-SA 3.0 IGO.

[39] M. Wang, R. Y. Kim, M. R. J. Kohonen-Corish *et al*., "Particulate matter air pollution as a cause of lung cancer: epidemiological and experimental evidence," *Br. J. Cancer*, vol. 132, pp. 986–996, 2025, doi: 10.1038/s41416-025-02999-2.

[40] D. Li, J. Shi, D. Liang, M. Ren, and Y. He, "Lung cancer risk and exposure to air pollution: a multicenter North China case-control study involving 14604 subjects," *BMC Pulmonary Medicine*, vol. 23, no. 1, p. 182, 2023, doi: 10.1186/s12890-023-02480-x.

[41] S. Jun *et al*., "Cancer risk based on alcohol consumption levels: a comprehensive systematic review and meta-analysis," *Epidemiol Health*, vol. 45, e2023092, Oct. 2023, doi: 10.4178/epih.e2023092.

[42] C. Bertola, C. Gobbetti, G. Baccarini, and R. Fabiani, "Wine consumption and lung cancer risk: A systematic review and meta-analysis," *Nutrients*, vol. 17, no. 8, p. 1322, 2025, doi: 10.3390/nu17081322.

[43] S. Peters *et al*., "Occupational exposure to organic dust increases lung cancer risk in the general population," *Thorax*, vol. 67, no. 2, pp. 111–116, 2012, doi: 10.1136/thoraxjnl-2011-200716.

[44] D. Wang, W. Li, N. Albasha *et al*., "Long-term exposure to house dust mites accelerates lung cancer development in mice," *J. Exp. Clin. Cancer Res.*, vol. 42, p. 26, 2023, doi: 10.1186/s13046-022-02587-9.

[45] M. Pyambri, S. Lacorte, J. Jaumot, and C. Bedia, "Effects of indoor dust exposure on lung cells: Association of chemical composition with phenotypic and lipid changes in a 3D lung cancer cell model," *Ecotoxicol. Public Health*, Nov. 2023, doi: 10.1021/acs.est.3c07573.

[46] M. Xu *et al*., "Prevalent occupational exposures and risk of lung cancer among women: Results from the application of the Canadian Job-Exposure Matrix (CANJEM) to a combined set of ten case–control studies," *Am. J. Ind. Med.*, early access, Jan. 8, 2024, doi: 10.1002/ajim.23562.

[47] J. S. Thakur, A. Rana, R. Kaur, and S. Malhotra, "Exposure to occupational carcinogens and risk of lung cancer: A systematic review and meta-analysis," *Int. J. Noncommun. Dis.*, vol. 8, no. 3, pp. 129–136, Jul.–Sep. 2023, doi: 10.4103/jncd.jncd_50_23.

[48] T. C. García, A. Ruano-Ravina, C. Candal-Pedreira *et al*., "Occupation as a risk factor of small cell lung cancer," *Sci. Rep.*, vol. 13, p. 4727, 2023, doi: 10.1038/s41598-023-31991-0.

[49] H. Yuan, Y. Wang, and H. Duan, "Risk of lung cancer and occupational exposure to polycyclic aromatic hydrocarbons among workers cohorts — Worldwide, 1969–2022," *China CDC Weekly*, Apr. 29, 2022, doi: 10.46234/ccdcw2022.085.

[50] K. R. Starke, U. Bolm-Audorff, D. Reissig, and A. Seidler, "Dose-response-relationship between occupational exposure to diesel engine emissions and lung cancer risk: A systematic review and meta-analysis," *Int. J. Hyg. Environ. Health*, vol. 256, Art. no. 114299, Mar. 2024, doi: 10.1016/j.ijheh.2023.114299.

[51] B. Kim, E. Y. Park, J. Kim, E. Park, J. K. Oh, and M. K. Lim, "Occupational exposure to pesticides and lung cancer risk: A propensity score analyses," *Cancer Res. Treat.*, vol. 54, no. 1, pp. 130–139, 2022, doi: 10.4143/crt.2020.1106.

[52] L. Ang, C. P. Y. Chan, W. P. Yau, and W. J. Seow, "Association between family history of lung cancer and lung cancer risk: A systematic review and meta-analysis," *Lung Cancer*, vol. 148, pp. 129–137, 2020, doi: 10.1016/j.lungcan.2020.08.012.

[53] F. Citarella *et al*., "Clinical implications of the family history in patients with lung cancer: A systematic review of the literature and a new cross-sectional/prospective study design (FAHIC: lung)," *J. Transl. Med.*, vol. 22, p. 714, 2024, doi: 10.1186/s12967-024-05538-4.

[54] W. Wei, S. Wang, Z. Yuan, *et al*., "Plant-based diets and the risk of lung cancer: a large prospective cohort study," *European Journal of Nutrition*, vol. 64, p. 73, 2025. doi: 10.1007/s00394-024-03570-0.

Vol. 6, No. 4, Agustus 2025, Page. 2187-2206
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4925

[55] H. Yan, X. Jin, C. Zhang, *et al*., "Associations between diet and incidence risk of lung cancer: A Mendelian randomization study," *Frontiers in Nutrition*, vol. 10, Mar. 2023. doi: 10.3389/fnut.2023.1149317.

[56] L. Peng, Q. Du, L. Xiang, *et al*., "Adherence to the low-fat diet pattern reduces the risk of lung cancer in American adults aged 55 years and above: a prospective cohort study," *The Journal of Nutrition, Health and Aging*, vol. 28, no. 7, p. 100240, Jul. 2024. doi: 10.1016/j.jnha.2024.100240.

[57] H. Wu, J. Yang, H. Wang, and L. Li, "Mendelian randomization to explore the direct or mediating associations between socioeconomic status and lung cancer," *Front. Oncol.*, vol. 13, Art. no. 1143059, 2023, doi: 10.3389/fonc.2023.1143059.

[58] C. Faselis *et al*., "Assessment of lung cancer risk among smokers for whom annual screening is not recommended," *JAMA Oncol.*, vol. 8, no. 10, pp. 1428–1437, 2022, doi: 10.1001/jamaoncol.2022.2952.

[59] D. S. Gutiérrez-Torres *et al*., "Changes in smoking use and subsequent lung cancer risk in the Alpha-Tocopherol, Beta-Carotene Cancer Prevention Study," *J. Natl. Cancer Inst.*, vol. 116, no. 6, pp. 895–901, 2024, doi: 10.1093/jnci/djae012.

[60] F. Lotfi *et al*., "Exposure to second-hand smoke and risk of lung cancer among Iranian population: A multicenter case-control study," *PLOS ONE*, Jul. 10, 2024, doi: 10.1371/journal.pone.0306517.

[61] S. Elkefi, G. Zeinoun, A. Tounsi, J. M. Bruzzese, C. Lelutiu-Weinberger, and A. K. Matthews, "Second-hand smoke exposure and risk of lung cancer among nonsmokers in the United States: A systematic review and meta-analysis," *Int. J. Environ. Res. Public Health*, vol. 22, no. 4, p. 595, 2025, doi: 10.3390/ijerph22040595.

[62] L. M. S. Sætre, K. Balasubramaniam, J. Søndergaard *et al*., "Smoking status, symptom significance and healthcare seeking with lung cancer symptoms in the Danish general population," *npj Prim. Care Respir. Med.*, vol. 35, p. 3, 2025, doi: 10.1038/s41533-025-00412-2.

[63] S. Zhang, Y. Deng, X. Xiang, Q. Xu, L. Hu, M. Xia, and L. Liu, "Postoperative symptom network analysis in non-small cell lung cancer patients: A cross-sectional study," *BMC Pulm. Med.*, vol. 25, no. 1, p. 244, 2025, doi: 10.1186/s12890-025-03711-z

[64] B. C. Bade *et al*., "Cancer-related fatigue in lung cancer: A research agenda: An official American Thoracic Society research statement," *Am. J. Respir. Crit. Care Med.*, vol. 207, no. 5, 2023, doi: 10.1164/rccm.202210-1963ST.

[65] J. Shin *et al*., "Distinct shortness of breath profiles in oncology outpatients undergoing chemotherapy," *J. Pain Symptom Manage.*, vol. 65, no. 3, pp. 242–255, Mar. 2023, doi: 10.1016/j.jpainsymman.2022.11.010.

[66] A. Qdaisat *et al*., "Severity of symptoms as an independent predictor of poor outcomes in patients with advanced cancer presenting to the emergency department: Secondary analysis of a prospective randomized study," *Cancers*, vol. 16, no. 23, p. 3988, 2024, doi: 10.3390/cancers16233988.

[67] S. Marmor, S. Cohen, N. Fujioka, L. C. Cho, A. Bhargava, and S. Misono, "Dysphagia prevalence and associated survival differences in older patients with lung cancer: A SEER-Medicare population-based study," *J. Geriatr. Oncol.*, vol. 11, no. 7, pp. 1115–1117, 2020, doi: 10.1016/j.jgo.2020.02.015.

[68] P. Obarski and J. Włodarczyk, "Alleviation of malignant dysphagia in inoperable lung cancer," *Ann. Palliat. Med.*, vol. 12, no. 4, pp. 738–747, 2023, doi: 10.21037/apm-22-1144.

[69] K. M. Udayappan and C. V. Anstine, "What diagnostic tests should be done after discovering clubbing in a patient without cardiopulmonary symptoms?," *Cleve. Clin. J. Med.*, vol. 92, no. 5, pp. 273–276, May 2025, doi: 10.3949/ccjm.92a.24052.

[70] J. Zheng et al., "Hospital-treated infectious diseases, infection burden, and risk of lung cancer: An observational and Mendelian randomization study," *Chest*, vol. 167, no. 1, pp. 270–282, Jan. 2025, doi: 10.1016/j.chest.2024.06.3811.

[71] L. Shahkar, N. Bigdeli, M. Mazandarani, and N. Lashkarbolouk, "A rare case of pulmonary adenocarcinoma in an 8-year-old patient with persistent respiratory manifestation: A case report study," *Case Rep. Oncol.*, vol. 16, no. 1, pp. 739–745, Aug. 2023, doi: 10.1159/000531986. PMID: 37933310.

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 4, Agustus 2025, Page. 2187-2206
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4925