Report: Lung Cancer Prediction Project with AI Application

July 20, 2025

Contents

1 Project Objective and Key Results

1.1 Project Objective

The primary goal of this project is to analyze patient health data to identify key risk factors associated with lung cancer and to develop an AI-powered platform capable of predicting individual risk levels. By leveraging data analytics and machine learning techniques, the project aims to enhance early detection and prevention of this life-threatening disease. Through the identification of behavioral and environmental patterns, the system can offer data-driven insights and support for both individuals and healthcare professionals, contributing to improved public health outcomes and more efficient medical interventions.

1.2 Key Results

- **Data Acquisition & Preprocessing:** We successfully collected and preprocessed a dataset containing health-related records of **1,000 patients**. The dataset includes **21 features** covering behavioral habits (e.g., smoking, alcohol use), environmental exposures (e.g., air pollution, occupational hazards), genetic factors, and clinical symptoms (e.g., fatigue, chest pain, coughing blood).

- **Exploratory Data Analysis (EDA):** A comprehensive analysis was conducted to explore data distribution, identify trends, and understand correlations between variables. This step helped to uncover potential links between certain lifestyle factors and the risk of developing lung cancer.

- **Predictive Modeling:** Machine learning models were trained using the processed data to predict lung cancer risk. These models achieved a prediction accuracy of **95%** *(to be updated once evaluation metrics are finalized)*. The models not only provide risk predictions but also highlight which features contribute most significantly to the outcome.

- **Chatbot Deployment:** An interactive chatbot was developed using the **Botpress** platform. The chatbot is capable of answering questions related to lung cancer, assessing risk based on user input, and providing educational support. This tool helps users better understand their health status and the potential risk factors associated with lung cancer.

- **AI-Powered Web Application:** A web-based application was developed using **Streamlit**. This application integrates the entire AI pipeline—from data ingestion and preprocessing, to prediction and user interaction. It serves as a complete decision-support system, offering risk evaluations and chatbot consultations in a seamless interface.

2 Business Understanding

2.1 Problem Statement

Lung cancer remains one of the world's deadliest cancers and a leading cause of cancer-related deaths. In 2020, approximately 2.2 million new cases were diagnosed, with 1.8 million resulting in death. A critical challenge lies in early detection, since symptoms in initial stages are often absent or extremely subtle—such as persistent fatigue, slight changes in voice, or finger clubbing—making them easy to overlook. Consequently, most patients are diagnosed at advanced stages, where five-year survival rates can drop below 10%, compared to over 70% if detected at stage I.

Modern screening techniques like low-dose CT (LDCT) scans have demonstrated a 20–25% reduction in mortality among high-risk individuals by enabling earlier detection. However, barriers such as limited access, low awareness, and patient hesitance cause less than 5% of eligible populations worldwide to undergo regular screening. Improving early detection remains a pressing need.

Given this context, our research aims to answer the following question:

> *How can we leverage routine health and behavioral data to predict lung cancer risk early—before symptoms emerge or imaging is performed?*

2.2 Value Proposition and Stakeholder Benefits

For Individuals

- **Proactive health monitoring:** By inputting their data (e.g., smoking habits, environmental exposures, symptom indicators), users can assess their risk level.
- **Empowerment to act:** Early prediction may prompt users to seek medical evaluation or adjust lifestyle choices before severe illness develops.

For Healthcare Professionals

- **Insight into key risk drivers:** Feature importance analysis will highlight the most influential risk factors (e.g., smoking, air pollution, genetic predispositions), helping professionals prioritize interventions.
- **Prioritized care delivery:** Identifying high-risk individuals enables targeted follow-up, such as LDCT screening or specialist referrals.

For Healthcare Systems

- **Reduced clinician workload:** A chatbot that handles initial risk assessments and provides educational support can alleviate pressure on healthcare teams.
- **Efficient preventative care:** Early detection models can help shift resources from late-stage treatment to early-stage screening, improving health outcomes and reducing cost burden.

2.3 Strategic Alignment

This project aligns with key priorities in modern healthcare innovation and digital transformation. Specifically, it supports the following strategic objectives:

- **Improved Public Health Outcomes:** By promoting early detection of lung cancer through AI-based risk prediction, the system contributes to timely medical intervention, potentially reducing mortality and improving prognosis for high-risk individuals.

- **Healthcare Digitalization:** The integration of machine learning models, chatbot interfaces, and a web-based application reflects the ongoing shift toward intelligent, data-driven digital health systems. The use of open-source tools such as `Flask`, `Botpress`, and `Scikit-learn` exemplifies scalable and cost-effective innovation.

- **Patient Empowerment and Engagement:** The platform provides users with accessible, personalized risk assessment tools that enable self-awareness and proactive health decisions. Through conversational AI, individuals can explore their risk profiles, ask relevant health questions, and receive preventive guidance without needing immediate access to clinical professionals.

- **Scalability and Health System Support:** The chatbot's ability to automate initial triage and deliver basic health education alleviates strain on clinical resources, making it a valuable complement to traditional screening workflows, especially in resource-limited or underserved settings.

3 Data Understanding

3.1 Dataset Overview

The dataset used in this project is a comprehensive collection of patient information related to lung cancer risk. It contains **1,000 records**, each representing a unique patient, along with **26 columns** detailing demographic data, lifestyle habits, environmental exposures, genetic predispositions, clinical symptoms, and the final diagnosis or risk level.

This dataset is specifically designed to support analysis of potential causes and indicators of lung cancer, thereby facilitating early risk prediction through data science and AI techniques.

3.2 Data Source and Context

The dataset includes patient health and behavioral data such as:

- **Demographics**: e.g., Age, Gender
- **Environmental factors**: e.g., Air Pollution, Occupational Hazards
- **Lifestyle behaviors**: e.g., Alcohol Use, Smoking, Passive Smoker, Balanced Diet
- **Genetic and chronic conditions**: e.g., Genetic Risk, Chronic Lung Disease, Obesity
- **Symptoms and clinical signs**: e.g., Coughing of Blood, Chest Pain, Shortness of Breath, Clubbing of Finger Nails

According to a study published in *Nature Medicine*, air pollution can increase the risk of lung cancer even in nonsmokers. Over 462,000 people in China were studied for 6 years, revealing that those living in areas with higher pollution levels were more likely to develop lung cancer—especially among nonsmokers and older individuals. This highlights the importance of analyzing a wide array of non-traditional risk factors beyond smoking alone.

3.3 Data Structure

`df.info()` yields:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 26 columns):
 #   Column                      Non-Null Count  Dtype
 ---  ------                      --------------  -----
 0    index                       1000 non-null   int64
 1    Patient Id                  1000 non-null   object
 2    Age                         1000 non-null   int64
 3    Gender                      1000 non-null   int64
 4    Air Pollution               1000 non-null   int64
 5    Alcohol use                1000 non-null   int64
 6    Dust Allergy               1000 non-null   int64
 7    OccuPational Hazards       1000 non-null   int64
 8    Genetic Risk               1000 non-null   int64
 9    chronic Lung Disease       1000 non-null   int64
 10   Balanced Diet              1000 non-null   int64
 11   Obesity                    1000 non-null   int64
 12   Smoking                    1000 non-null   int64
 13   Passive Smoker             1000 non-null   int64
 14   Chest Pain                 1000 non-null   int64
 15   Coughing of Blood          1000 non-null   int64
 16   Fatigue                    1000 non-null   int64
 17   Weight Loss                1000 non-null   int64
 18   Shortness of Breath        1000 non-null   int64
 19   Wheezing                   1000 non-null   int64
 20   Swallowing Difficulty      1000 non-null   int64
 21   Clubbing of Finger Nails   1000 non-null   int64
 22   Frequent Cold              1000 non-null   int64
 23   Dry Cough                  1000 non-null   int64
 24   Snoring                    1000 non-null   int64
 25   Level                      1000 non-null   object
```

Data size: 1,000 rows × 26 columns

Missing values: None

Data types:

- 24 columns are of `int64` type
- 2 columns are of `object` type — `Patient Id` and `Level` (target)

3.4 Key Variables

Variable	**Type**	**Description**
Age	Numeric	Age of the patient
Gender	Categorical	0 = Male, 1 = Female
Air Pollution	Ordinal	Exposure level to polluted air
Alcohol use	Ordinal	Frequency or quantity of alcohol consumption
Dust Allergy	Ordinal	Allergy sensitivity to dust
Occupational Hazards	Ordinal	Exposure to workplace-related health risks
Genetic Risk	Ordinal	Genetic predisposition or family history of lung cancer
Chronic Lung Disease	Binary	Presence of chronic lung diseases (e.g., COPD, asthma)
Balanced Diet	Ordinal	Quality or consistency of maintaining a balanced diet
Obesity	Ordinal	Obesity level or BMI-based classification
Smoking, Passive Smoker	Ordinal	Direct and secondhand smoking exposure
Chest Pain, Fatigue, Coughing of Blood, etc.	Binary	Symptom presence (0 = No, 1 = Yes)
Level	Categorical	**Target variable** indicating lung cancer risk level

Table 1: Key variables and their types

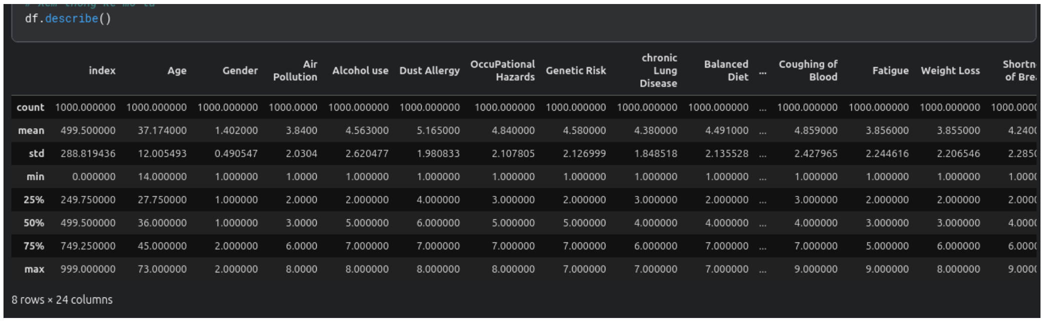
3.5 Preliminary Observations

- All features are **complete** with no missing values.
- Most variables are **categorical or ordinal**, requiring appropriate encoding (e.g., label encoding, one-hot encoding).

- The target variable `Level` should be analyzed for class distribution to assess class imbalance.
- Potentially correlated features include `Smoking` and `Passive Smoker`, or `Fatigue` and `Weight Loss`.

3.6 Descriptive Statistics

A statistical summary of the numerical and ordinal features provides insight into central tendencies, variability, and range of values in the dataset. The summary is based on `df.describe()` for the 24 quantitative columns (excluding `index` and `Patient Id`).



```
df.describe()
```

	index	Age	Gender	Air Pollution	Alcohol use	Dust Allergy	OccuPational Hazards	Genetic Risk	chronic Lung Disease	Balanced Diet	...	Coughing of Blood	Fatigue	Weight Loss	Shortn of Bre
count	1000.000000	1000.000000	1000.000000	1000.0000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	...	1000.000000	1000.000000	1000.000000	1000.0000
mean	499.500000	37.174000	1.402000	3.8400	4.563000	5.165000	4.840000	4.580000	4.380000	4.491000	...	4.859000	3.856000	3.855000	4.240
std	288.819436	12.005493	0.490547	2.0304	2.620477	1.980833	2.107805	2.126999	1.848518	2.135528	...	2.427965	2.244616	2.206546	2.285
min	0.000000	14.000000	1.000000	1.0000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	...	1.000000	1.000000	1.000000	1.000
25%	249.750000	27.750000	1.000000	2.0000	2.000000	4.000000	3.000000	2.000000	3.000000	2.000000	...	3.000000	2.000000	2.000000	2.000
50%	499.500000	36.000000	1.000000	3.0000	5.000000	6.000000	5.000000	5.000000	4.000000	4.000000	...	4.000000	3.000000	3.000000	4.000
75%	749.250000	45.000000	2.000000	6.0000	7.000000	7.000000	7.000000	7.000000	6.000000	7.000000	...	7.000000	5.000000	6.000000	6.000
max	999.000000	73.000000	2.000000	8.0000	8.000000	8.000000	8.000000	7.000000	7.000000	7.000000	...	9.000000	9.000000	8.000000	9.000

8 rows × 24 columns

Figure 1: Dataframe Describe

- **Count:** All features contain **1,000 non-null** entries, confirming the absence of missing values.
- **Age:**
  - Mean: 37.17 years Median: 36 years
  - Standard Deviation: 12.01 years
  - Range: 14 to 73 years
  - Interquartile Range (IQR): 27.75 to 45 years

  The age distribution suggests a relatively young cohort, with 50% of patients between their late 20s and mid-40s.

- **Gender:** (encoded as integers, likely 1 = Male, 2 = Female)

  - Mean: 1.402   Std Dev: 0.491   Min–Max: 1–2

  The mean indicates a higher proportion of one gender—likely males (60%)—in the sample.

- **Environmental and Lifestyle Factors:**
  These are encoded on ordinal scales (typically 1–7, 1–8, or 1–9), representing exposure or behavioral intensity.

  - Air Pollution: Mean = 3.84   Range: 1–8   IQR: 2–6
  - Alcohol Use: Mean = 4.56   Range: 1–8   IQR: 2–7
  - Dust Allergy: Mean = 5.17   Range: 1–8   IQR: 4–7
  - Occupational Hazards: Mean = 4.84   Range: 1–8   IQR: 3–7
  - Genetic Risk: Mean = 4.58   Range: 1–7   IQR: 2–7
  - Chronic Lung Disease: Mean = 4.38   Range: 1–7   IQR: 3–6
  - Balanced Diet: Mean = 4.49   Range: 1–7   IQR: 2–7
  - Obesity: Mean = 4.52   (similar distribution to above)

  These features show moderate average risk/exposure levels and substantial variability (standard deviations ~ 2.0), suggesting a heterogeneous cohort.

- **Smoking and Passive Smoking:**

  - Smoking: Mean = 4.86   Range: 1–9   IQR: 3–7
  - Passive Smoker: Mean = 3.86   Range: 1–8   IQR: 2–5

  Direct smoking is more prevalent than passive smoking, with several patients reporting high levels of smoking behavior.

- **Clinical Symptoms:** These binary or ordinal indicators reflect the presence or severity of symptoms.

  - Common symptoms such as `Chest Pain`, `Fatigue`, `Weight Loss`, `Shortness of Breath`, and `Coughing of Blood` exhibit means in the range of 3.75 to 4.86.
  - Less frequent symptoms include:
    - Frequent Cold: Mean = 3.54 Range: 1–7 IQR: 2–5
    - Dry Cough: Mean = 3.85 Range: 1–7 IQR: 2–6
    - Snoring: Mean = 2.93 Range: 1–7 IQR: 2–4

- **Variability and Outliers:**
  - Many features have standard deviations near 2.0, indicating moderate-to-high spread.
  - Minimum values are typically 1; maximums vary by scale (up to 9 in some cases), suggesting potential outliers.

4 Exploratory Data Analysis

4.1 Alcohol Use vs. Disease Level

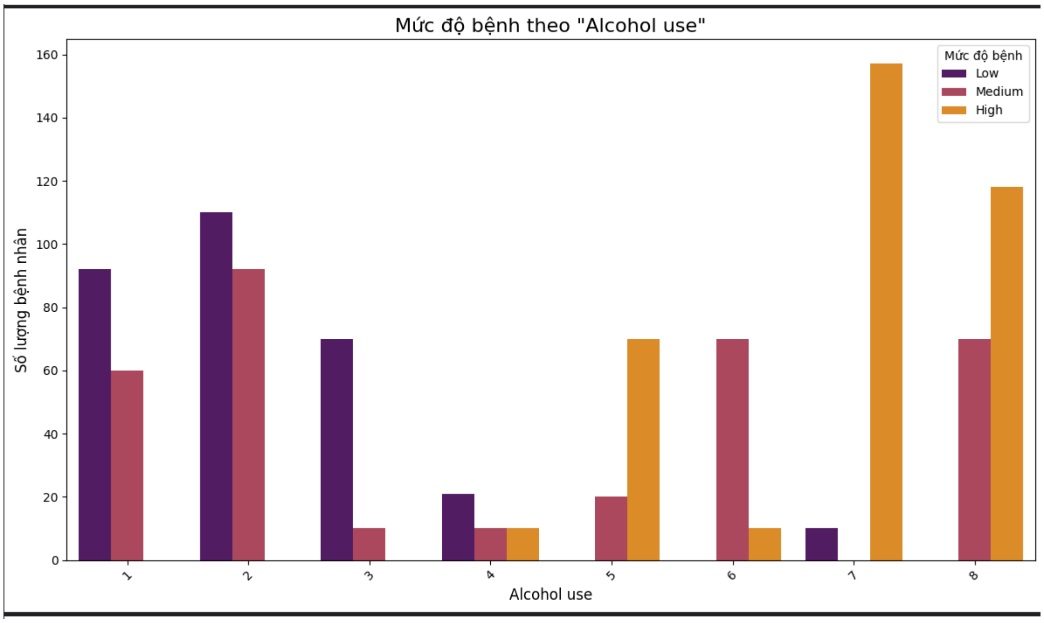


Figure 2: Alcohol Use vs. Lung Cancer Risk Level

The relationship between alcohol consumption and lung cancer risk reveals a clear pattern of increasing severity with higher reported usage. At the lower end of the alcohol use spectrum (levels 1 to 3), patients predominantly fall into the Low-risk category. For instance, at level 1, more than 90 patients are classified as Low risk, compared to only around 60 at Medium risk and virtually none at High. This trend continues through level 3.

As alcohol use increases to moderate levels (levels 4 to 6), the distribution shifts. Level 5 marks a turning point where Medium risk surpasses Low, and High-risk patients begin to appear in greater numbers. At level 6, Medium risk peaks while High-risk patients begin increasing.

The most significant transformation occurs at levels 7 and 8. Level 7 shows over 160 individuals classified as High risk, with very few Low or Medium. At level 8, the pattern remains consistent with High risk dominating. This trend supports the hypothesis that increased alcohol consumption is positively correlated with lung cancer susceptibility.

4.2 Dust Allergy vs. Disease Level

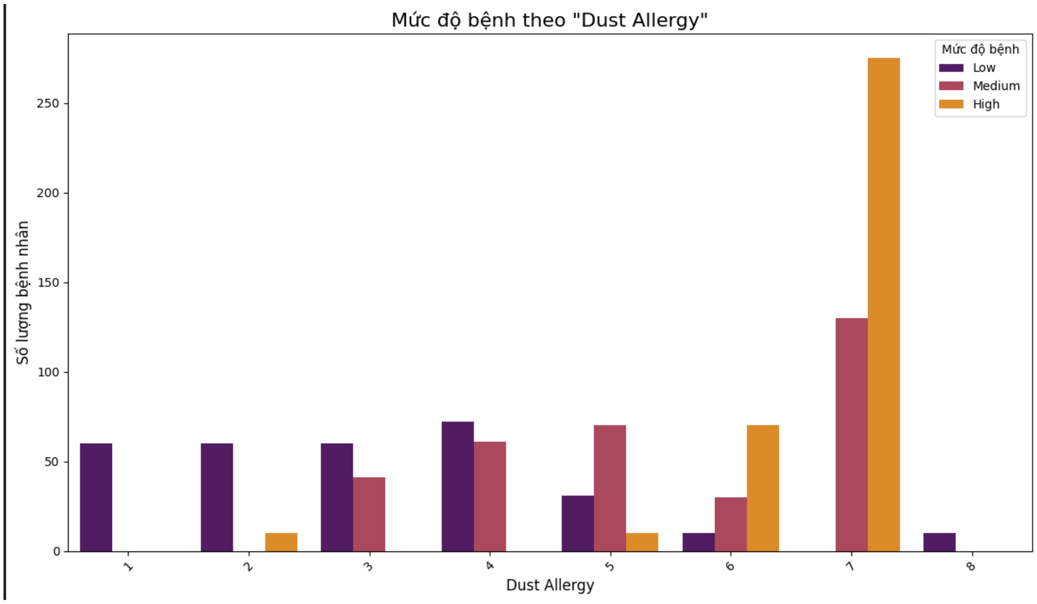


Figure 3: Dust Allergy vs. Lung Cancer Risk Level

Dust allergy severity also correlates strongly with lung cancer risk. Levels 1 through 3 are largely associated with Low-risk patients. As dust allergy levels reach 4 and 5, Medium-risk patients increase notably, and High-risk begins to appear.

At higher allergy levels, particularly level 6, High-risk classifications become dominant. Level 7 exhibits a dramatic escalation with the vast majority of patients assigned to High risk. Interestingly, at level 8, Medium risk dominates while Low and High vanish, possibly indicating an edge case or data sparsity. These results imply a threshold effect near levels 5–6, where dust allergy severity becomes a significant predictive factor.

4.3 Occupational Hazards vs. Disease Level

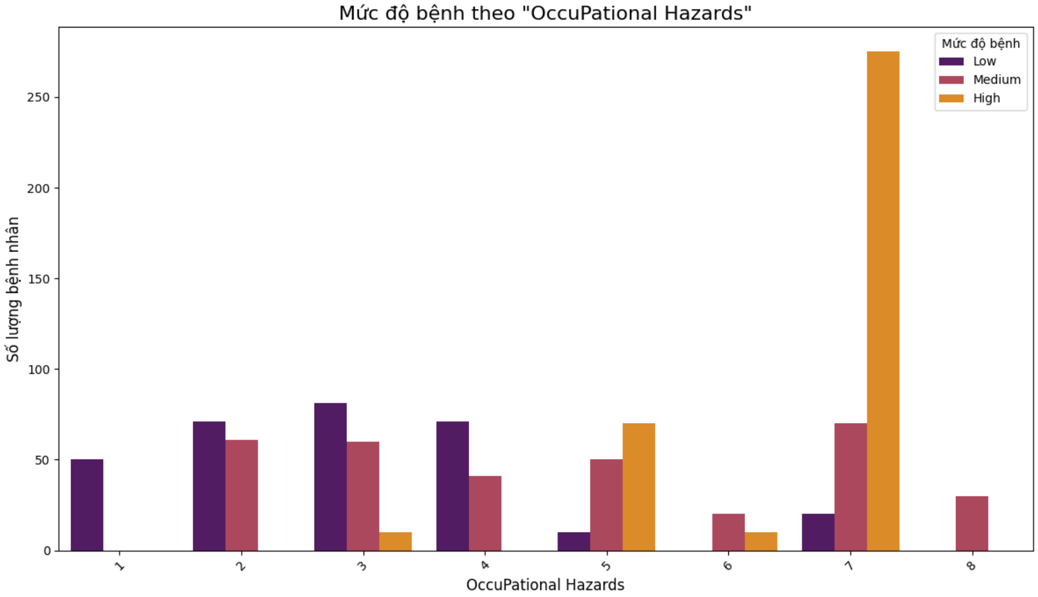


Figure 4: Occupational Hazards vs. Lung Cancer Risk Level

For lower exposure levels (1–3), most patients fall into the Low-risk group, with very few classified as Medium or High risk. As exposure increases into moderate levels (4–5), Medium risk rises, and High risk becomes more prominent. Level 5 in particular shows a strong presence of High-risk individuals.

Levels 6 to 8 reveal a major shift toward High-risk classification, especially at level 7 where High-risk patients form the majority. Level 8 has only Medium-risk entries, potentially due to limited sample size. These patterns suggest that occupational hazard exposure beyond level 4 significantly raises cancer risk.

4.4 Balanced Diet vs. Disease Level

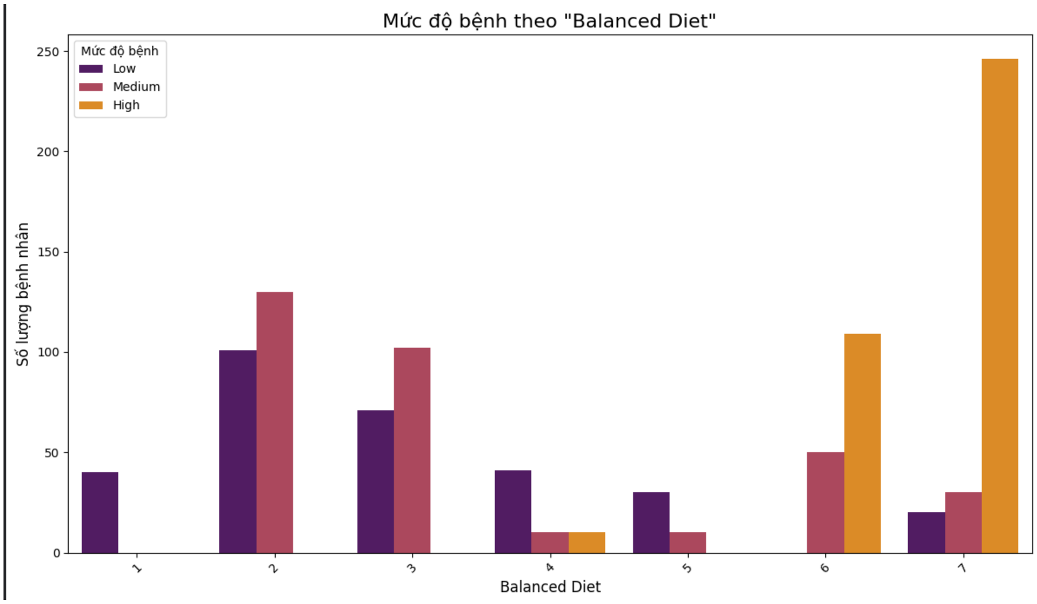


Figure 5: Balanced Diet vs. Lung Cancer Risk Level

The relationship between balanced diet and disease level is non-monotonic. Levels 1 to 3 contain a mix of Low and Medium-risk patients, with few High-risk cases. At levels 4 and 5, the presence of Low-risk patients increases, indicating some protective association.

Paradoxically, levels 6 and 7, which indicate the highest reported diet quality, are dominated by High-risk classifications. For example, at level 7, nearly all patients are High risk. This counterintuitive trend may reflect confounding effects, such as patients adopting better diets due to prior illness, or potential reporting inconsistencies. Therefore, dietary score alone may not predict cancer risk reliably without accounting for other covariates.

4.5 Chronic Lung Disease vs. Disease Level

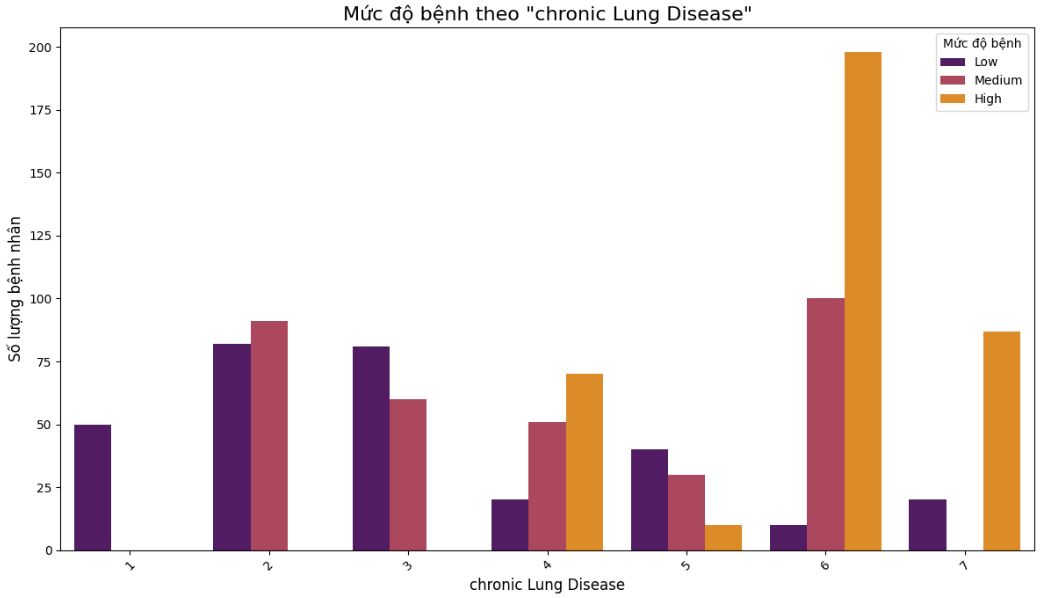


Figure 6: Chronic Lung Disease vs. Lung Cancer Risk Level

Chronic lung disease severity is a potent and consistent predictor of lung cancer risk. At levels 1 to 3, the patient population is primarily categorized as Low or Medium risk, with minimal High-risk cases.

At levels 4 and 5, Medium and High-risk cases rise significantly, surpassing Low risk. The trend becomes especially pronounced at levels 6 and 7, where the vast majority of patients are classified as High risk. At level 7, all patients fall into the High-risk group, reinforcing clinical findings that chronic lung conditions are a major driver of lung cancer risk.

4.6 Distribution of Genetic Risk by Cancer Level

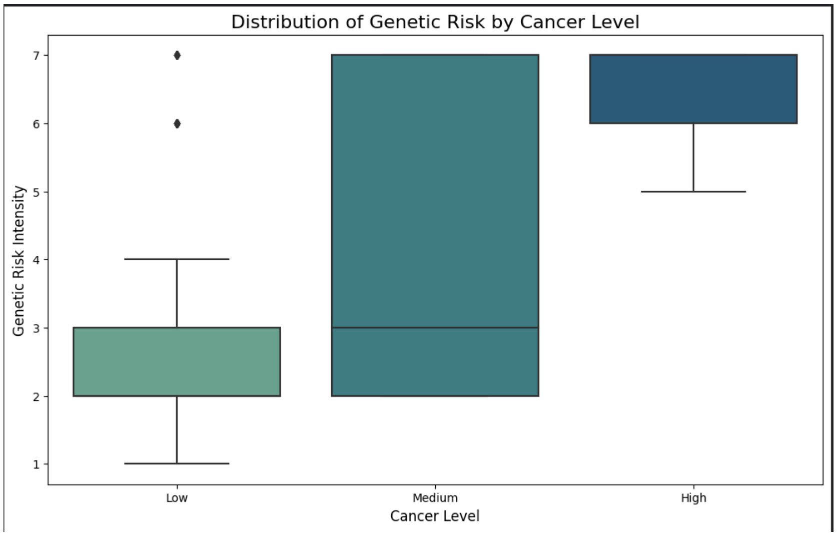


Figure 7: Boxplot of Genetic Risk across Lung Cancer Severity Levels

The boxplot above illustrates the statistical distribution of the **Genetic Risk** variable across three categories of lung cancer severity: **Low**, **Medium**, and **High**. Each box encapsulates the interquartile range (IQR) from the 25th percentile (Q1) to the 75th percentile (Q3), with the central line denoting the median. Points outside the whiskers indicate statistical outliers.

Key Observations:

- **Low-risk patients** exhibit the lowest genetic predisposition, with a median value around **2.5** and most observations falling within the range of **2 to 4**. A few outliers extend as high as 6–7, but these are rare.

- **Medium-risk individuals** display a broader distribution of genetic risk scores. The interquartile spread ranges from approximately **2 to 7**, with a median centered near **3**, suggesting considerable heterogeneity in genetic risk within this group.

- **High-risk patients** exhibit a remarkably narrow and elevated distribution. The genetic risk scores in this group are concentrated at the upper bound (values **6–7**), reflecting high consistency in genetic vulnerability among those most severely affected.

Interpretation: Genetic Risk demonstrates a **positive monotonic association** with lung cancer severity. As the risk level increases from Low to High, both the median and concentration of genetic risk escalate significantly. Moreover, the homogeneity in the High group suggests that this variable could serve as a **strong discriminative feature** in predictive modeling.

4.7 Distribution of Chronic Lung Disease by Cancer Level

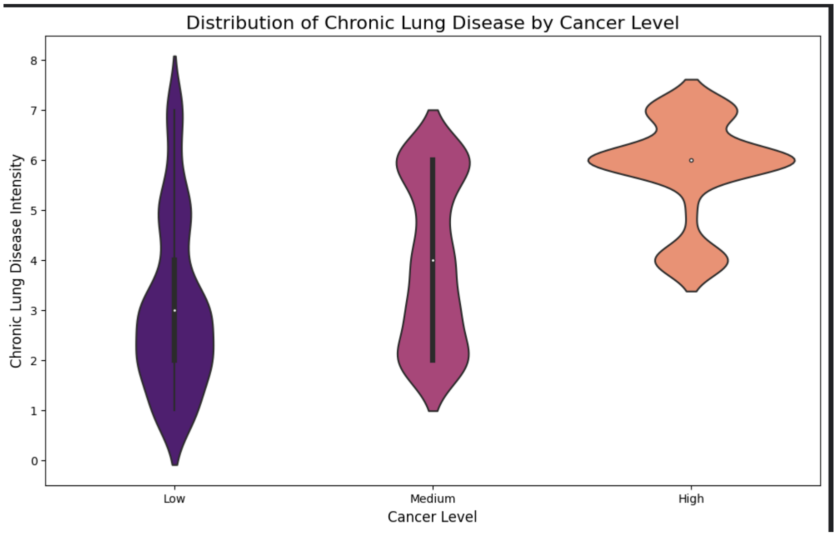


Figure 8: Violin Plot of Chronic Lung Disease Across Risk Groups

The violin plot presents the kernel density distribution of **chronic lung disease** severity across the same three cancer risk categories. In addition to depicting median and quartiles (via embedded boxplot), the violin plot reveals the **probability density** of values, enabling a richer understanding of data spread and concentration.

Key Observations:

- In the **Low-risk group**, the distribution is relatively uniform, with the highest density around **2–3**. The median lies near **3**, and the presence of severe chronic lung disease (values above 6) is rare.
- In the **Medium-risk category**, the distribution begins to shift. The density peaks around **4–6**, and the median shifts to approximately **4**, suggesting that chronic lung conditions become increasingly prevalent.
- The **High-risk group** displays a sharp and narrow density centered between **6 and 7**, with very little spread. This reflects a **near-universal prevalence of severe chronic lung disease** among individuals classified as high risk.

Interpretation: Chronic lung disease acts as a **transitional biomarker**, showing a clear, stepwise escalation as cancer severity increases. From a pathophysiological perspective, this supports the hypothesis that chronic inflammation and airway damage—common in long-standing lung diseases—may act as precursors or co-factors in cancer development.

4.8 Correlation Analysis

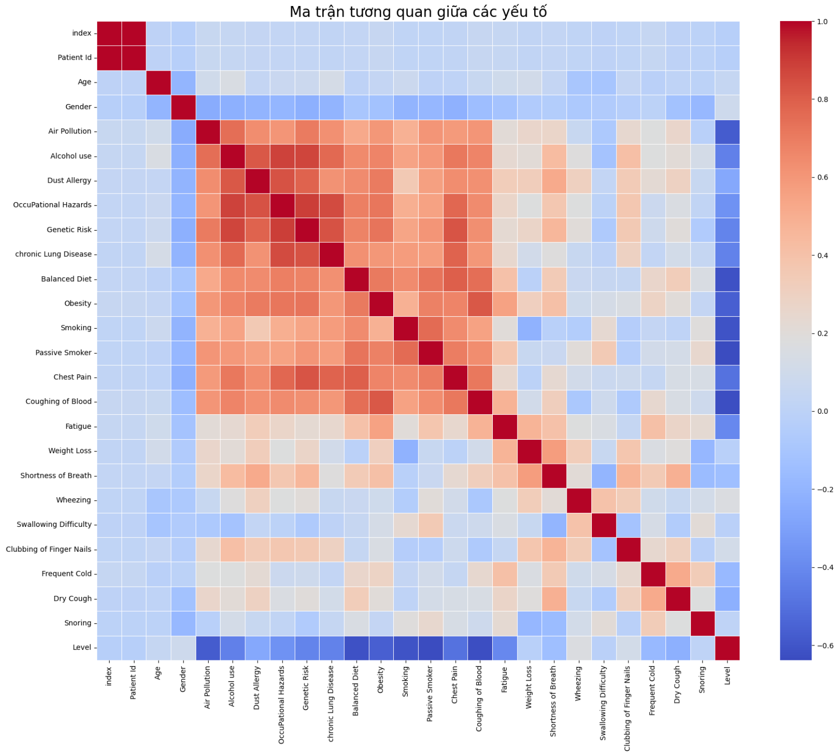


Figure 9: Pearson Correlation Heatmap between Features

To better understand the relationships between variables in the lung cancer dataset, a Pearson correlation matrix was computed and visualized using a heatmap (Figure above). All categorical variables were ordinally encoded to permit numerical correlation calculation. Several important patterns emerge from this analysis, offering both methodological and clinical insights.

A broad positive correlation cluster is observed among clinical symptom variables (e.g., `Chest Pain`, `Coughing of Blood`, `Fatigue`, `Shortness of Breath`, `Weight Loss`) and behavioral or environmental exposure factors such as `Smoking`, `Alcohol use`, `Air Pollution`, `Dust Allergy`, and `Occupational Hazards`. Correlation coefficients between these features often exceed 0.6, suggesting a latent severity axis where patients simultaneously accumulate risk behaviors and manifest symptoms. This reinforces the clinical intuition that lung cancer rarely emerges from a single isolated cause but rather from the interaction of multiple contributing factors. From a modeling standpoint, such redundancy can be beneficial for tree-based models like Random Forests or XGBoost, which can leverage shared variance. However, generalized linear models (e.g., logistic regression) may suffer from multicollinearity unless regularized.

Notably, certain lifestyle-related features exhibit very strong intercorrelations. For instance, `Smoking` and `Passive Smoker` show a correlation above 0.7, indicating that individuals who smoke often expose others around them, including cohabitants or coworkers, to secondhand smoke. Similarly, `Alcohol use` aligns closely with both `Smoking` and environmental factors like `Air Pollution` and `Occupational Hazards`. These features likely represent overlapping socio-economic or geographic effects and may collectively define a "lifestyle-exposure" construct. Dimensionality reduction techniques such as principal component analysis (PCA) may help consolidate these into unified explanatory variables.

Interestingly, `Balanced Diet`—typically viewed as a protective factor—shows a similar correlation profile to high-risk predictors. Higher values for this variable correlate positively with `Smoking`, `Air Pollution`, and symptom-related variables. This counterintuitive result may arise from incorrect labeling (e.g., higher numeric values representing poorer diets rather than better ones) or from reverse causality, where patients suffering from illness are more likely to adopt stricter diets after diagnosis. Further inspection and potential recoding of this variable are warranted before it is used in predictive modeling.

The variable `Chronic Lung Disease` emerges as a central "bridge" feature between lifestyle exposures and clinical outcomes. It is moderately correlated with upstream factors such as `Smoking` and `Dust Allergy`, and strongly correlated with downstream symptoms like `Coughing of Blood` and `Shortness of Breath`. In a path analysis framework, this suggests that chronic lung disease may serve as an intermediate variable—aggravated by long-term exposure and, in turn, amplifying cancer-related symptoms.

Correlation with the target variable, `Level`, also reveals important trends. Assuming the risk levels are encoded inversely (e.g., High risk = 1, Low risk = 3), the strong negative correlations observed between `Level` and most other predictors (especially symptoms and exposure variables) confirm their importance in classification. Correlation magnitudes often exceed $|\rho| > 0.6$, underscoring that these variables contain strong linear or monotonic signal. Because the target relationship is likely nonlinear and ordinal, models such as decision trees, ensemble methods, or ordinal regression may be better suited than simple linear classifiers.

Finally, multicollinearity is a significant consideration. Feature pairs like (`Air Pollution`, `Dust Allergy`) and (`Smoking`, `Passive Smoker`) exceed correlation thresholds of $r = 0.75$, which could distort coefficient estimates and model stability in regression-based approaches. Regularization techniques (e.g., Ridge or Lasso) or dimensionality reduction should be employed if such models are used. Alternatively, feature attribution methods like SHAP values can help tree-based models identify which features contribute the most to risk classification, despite collinearity.

In summary, the correlation matrix not only confirms clinical expectations regarding risk and symptom progression, but also guides preprocessing decisions and model selection strategies for the next phase of analysis.

5 Predictive Modeling

5.1 Theory and Application of Models

In this section, we describe the learning algorithms applied to the lung cancer risk classification task. Two powerful ensemble-based models are selected for their robustness and predictive capabilities: **Random Forest Classifier** and **Gradient Boosting Classifier** (including variants such as XGBoost and LightGBM). Both methods are tree-based but differ fundamentally in how they aggregate predictions.

5.1.1 Random Forest Classifier

Random Forest is an ensemble learning method based on the **bagging** (Bootstrap Aggregating) principle. It constructs multiple decision trees trained on random subsets (with replacement) of the original dataset and then combines their predictions through majority voting. By decorrelating trees and averaging their outputs, Random Forest reduces model variance and mitigates overfitting.

Mathematical Formulation. Given a dataset $D = \{(x_i, y_i)\}_{i=1}^{n}$, Random Forest generates T decision trees $\{h_t(x)\}_{t=1}^{T}$, each trained on a bootstrap sample $D_t \subset D$. For classification tasks, the aggregated output $\hat{y}$ is computed via majority vote:

$$\hat{y} = \text{mode}\left(\{h_t(x)\}_{t=1}^{T}\right) \tag{1}$$

At each tree split, only a random subset of m features (out of p total) is considered to further reduce correlation between individual trees.

Advantages.

- Handles high-dimensional data well.
- Resistant to overfitting compared to single decision trees.
- Provides feature importance for interpretability.

Implementation (Scikit-learn).

Listing 1: Random Forest using Scikit-learn

```python
from sklearn.ensemble import RandomForestClassifier

model_rf = RandomForestClassifier(
    n_estimators=100,
    max_depth=None,
    class_weight='balanced',
    random_state=42,
    n_jobs=-1
)
model_rf.fit(X_train, y_train)
```

Listing 2: Extract Feature Importances

```
importances = model_rf.feature_importances_
```

5.1.2 Gradient Boosting Classifier

Gradient Boosting is a **sequential** ensemble method that builds trees in a stage-wise manner. Each subsequent model is trained to predict the residual errors (negative gradients) of the previous model with respect to a given loss function. Unlike bagging, boosting focuses on reducing bias by correcting mistakes made by earlier learners.

Mathematical Formulation. A boosted model is expressed as:

$$F(x) = \sum_{t=1}^{T} \alpha_t h_t(x) \tag{2}$$

where:

- $h_t(x)$: weak learner (usually a shallow decision tree),
- α_t: learning rate or step size,
- $F(x)$: the overall model prediction.

The model is updated iteratively to minimize a loss function $\mathcal{L}(y, F(x))$, using gradient descent. For example, in logistic loss:

$$\mathcal{L}(y, F(x)) = -\sum_{i=1}^{n} \left[y_i \log p_i + (1 - y_i) \log(1 - p_i) \right] \tag{3}$$

At iteration t, the residuals are computed as:

$$r_i^{(t)} = -\left[\frac{\partial \mathcal{L}(y_i, F(x_i))}{\partial F(x_i)} \right] \tag{4}$$

These residuals serve as pseudo-targets for the next learner.

Advantages.

- Typically achieves higher accuracy than bagging methods.
- Optimizes directly for predictive performance via loss minimization.
- Highly customizable via hyperparameters and regularization.

Implementation.

Listing 3: GradientBoostingClassifier

```
from sklearn.ensemble import GradientBoostingClassifier

model_gb = GradientBoostingClassifier(
    n_estimators=100,
    learning_rate=0.1,
    max_depth=3,
    random_state=42
)
model_gb.fit(X_train, y_train)
```

Listing 4: XGBoost Classifier

```
from xgboost import XGBClassifier

model_xgb = XGBClassifier(
    n_estimators=100,
    learning_rate=0.1,
    max_depth=3,
    objective=’multi:softprob’,
    eval_metric=’mlogloss’,
    use_label_encoder=False
)
model_xgb.fit(X_train, y_train)
```

Listing 5: LightGBM Classifier

```
from lightgbm import LGBMClassifier

```

```
model_lgb = LGBMClassifier(
    n_estimators=100,
    learning_rate=0.1,
    objective='multiclass',
    random_state=42
)
model_lgb.fit(X_train, y_train)
```

5.2 Evaluation Metrics

To objectively assess model quality in the multi-class classification setting, we use a suite of standard evaluation metrics. These metrics offer different perspectives on performance, especially critical when class distribution is imbalanced (e.g., Low vs. High cancer risk).

5.2.1 Accuracy

Accuracy quantifies the proportion of correct predictions out of all predictions made:

$$\text{Accuracy} = \frac{1}{n}\sum_{i=1}^{n} 1(\hat{y}_i = y_i)$$

Where:

- $\hat{y}_i$ is the predicted label,
- y_i is the true label,
- $1(\cdot)$ is the indicator function,
- n is the total number of samples.

Scikit-learn code:

Listing 6: Accuracy Score

```
from sklearn.metrics import accuracy_score
accuracy = accuracy_score(y_test, y_pred)
```

5.2.2 Precision, Recall, and F1-Score

These are per-class metrics useful for understanding type I and type II errors, and are especially meaningful in healthcare contexts.

Precision:

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c}$$

Recall:

$$\text{Recall}_c = \frac{TP_c}{TP_c + FN_c}$$

F1-Score:

$$\text{F1}_c = 2 \cdot \frac{\text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c}$$

Where TP_c, FP_c, FN_c are true positives, false positives, and false negatives for class c.

Scikit-learn code:

Listing 7: Classification Report

```python
from sklearn.metrics import classification_report
print(classification_report(y_test, y_pred, target_names
   =['Low', 'Medium', 'High']))
```

Macro-averaged metrics:

```python
from sklearn.metrics import precision_score,
   recall_score, f1_score

precision = precision_score(y_test, y_pred, average='
   macro')
recall = recall_score(y_test, y_pred, average='macro')
f1 = f1_score(y_test, y_pred, average='macro')
```

5.2.3 Confusion Matrix

A confusion matrix $\mathbf{C} \in N^{k \times k}$ summarizes prediction performance by class. The entry C_{ij} counts the number of times class i was predicted as class j.

Interpretation:

- Diagonal elements: correct predictions
- Off-diagonal elements: misclassifications

Listing 8: Confusion Matrix Heatmap

```
from sklearn.metrics import confusion_matrix
import seaborn as sns
import matplotlib.pyplot as plt

cm = confusion_matrix(y_test, y_pred)
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues',
            xticklabels=class_names, yticklabels=
               class_names)
plt.xlabel('Predicted')
plt.ylabel('Actual')
plt.title('Confusion Matrix')
plt.show()
```

5.2.4 ROC Curve and AUC (Area Under Curve)

In multi-class classification, we use a One-vs-Rest (OvR) strategy to compute ROC curves and AUC scores. For each class c:

$$\text{TPR} = \frac{TP}{TP + FN}, \quad \text{FPR} = \frac{FP}{FP + TN}$$

AUC:

$$\text{AUC}_c = \int_0^1 \text{TPR}_c(\text{FPR})\, d\text{FPR}$$

Scikit-learn code:

Listing 9: ROC-AUC for Multi-class

```
from sklearn.preprocessing import label_binarize
from sklearn.metrics import roc_auc_score

# Binarize the labels for OvR scheme
y_test_bin = label_binarize(y_test, classes=[0, 1, 2])
```

```
y_pred_prob = model.predict_proba(X_test)

# Compute macro-average AUC
auc_score = roc_auc_score(y_test_bin, y_pred_prob,
                          average='macro', multi_class='
                             ovr')
```

Plot ROC Curves for Each Class:

```
from sklearn.metrics import roc_curve, auc

for i in range(3):  # for 3 classes
    fpr, tpr, _ = roc_curve(y_test_bin[:, i],
       y_pred_prob[:, i])
    plt.plot(fpr, tpr, label=f'Class {i} (AUC = {auc(fpr
       , tpr):.2f})')

plt.plot([0, 1], [0, 1], 'k--')
plt.xlabel("False Positive Rate")
plt.ylabel("True Positive Rate")
plt.title("ROC Curve (OvR)")
plt.legend()
plt.show()
```

Together, these metrics offer a multi-dimensional perspective on classification performance. Accuracy gives a global correctness measure, while Precision, Recall, and F1-Score provide class-specific insights. The Confusion Matrix highlights error patterns, and ROC-AUC curves evaluate threshold sensitivity and discrimination power. Such a thorough evaluation is crucial for healthcare AI applications like lung cancer risk prediction.

6 Chatbot Integration

In the evolving landscape of AI-powered healthcare, chatbots have emerged as a pivotal component in digital health ecosystems. Acting as intelligent, conversational agents, they facilitate real-time interaction between users and backend systems. In this project, the chatbot serves as a **frontline interface** that bridges end-users with the machine learning models used for lung cancer risk prediction. Its core function is to democratize access to personalized health insights, enabling users to assess their risk levels, understand contributing factors, and receive tailored recommendations—all through natural language interaction.

Figure 10: Chatbot

The chatbot collects user-provided data such as age, gender, symptom intensity, lifestyle behaviors (e.g., smoking, alcohol use), and environmental exposures. These inputs are processed and evaluated by a pre-trained machine learning classifier, which returns a predicted lung cancer risk level (e.g., `Low`, `Medium`, `High`). Results are presented in a user-friendly format, enhancing both **usability** and **early awareness**—two crucial pillars in cancer prevention and health behavior change.

6.1 What is a Chatbot?

A **chatbot** is a software system capable of simulating human conversation via textual or voice-based communication. The primary goal is to understand user inputs, identify underlying intent, and respond appropriately in a conversational format. Chatbots can be categorized by their underlying architecture:

- **Rule-based chatbots** operate on predefined decision trees or if-else logic.
- **Retrieval-based chatbots** select responses from a fixed response bank using keyword matching, pattern recognition, or heuristic ranking.
- **Generative chatbots**, powered by deep learning (e.g., Transformers), generate dynamic responses word-by-word, allowing for greater flexibility and contextualization.

In this project, we implement a **hybrid chatbot** that combines rule-based flows with AI-assisted modules. It is designed to answer health-related questions such as:

> *"What is my lung cancer risk if I smoke and experience shortness of breath?"*

and to guide users through a conversational questionnaire. Once the relevant features are gathered, the chatbot forwards this data to the predictive model. Based on the output, it returns a risk score along with an explanation and actionable suggestions for the user.

6.2 Benefits of Chatbot Integration

The inclusion of a chatbot in healthcare applications delivers numerous operational and clinical advantages:

- **Accessibility:** Chatbots operate continuously, offering 24/7 availability across devices such as smartphones, tablets, and web browsers. This is particularly important in remote or under-resourced settings.
- **Scalability:** Unlike human agents, chatbots can engage with thousands of users in parallel without performance degradation, making them ideal for public health outreach or mass screening campaigns.

- **Early Triage:** Chatbots collect structured health data and provide preliminary risk stratification using ML models. This assists users in making timely decisions about seeking medical attention.
- **Health Education:** The chatbot communicates lung cancer-related information in simple language, raising awareness about modifiable risk factors and preventative strategies.
- **System Efficiency:** By handling routine queries and basic assessments, chatbots reduce the cognitive load on healthcare providers, enabling clinicians to focus on more critical cases.
- **User Experience (UX):** Through natural, conversational interactions, chatbots promote trust, privacy, and personalization. Users are more likely to engage with AI systems that emulate empathetic dialogue.

6.3 How Many Types of Chatbots Are There?

Chatbots represent a spectrum of technologies with varying levels of intelligence, adaptability, and interaction design. Their classification depends on **how they process user input**, **how they decide on a response**, and **how they adapt to changing conversations**. Broadly, chatbots can be categorized into four main types, each with distinct advantages and limitations depending on the application context.

1. Rule-based Chatbots Rule-based chatbots are the most traditional and deterministic type. They operate on predefined logic, typically built using `if-else` conditions, decision trees, or button-driven menus. These bots follow scripted flows and only respond to inputs that match their predefined rules.

Use Cases:

- Appointment scheduling
- FAQ bots

- Preliminary health screening with checklists

Advantages:

- Simple to develop with no need for training data
- Predictable behavior suitable for regulated environments
- Low maintenance

Limitations:

- Inflexible to natural language variations
- No memory or context awareness
- Poor scalability as use cases grow

2. AI-powered Chatbots (Conversational AI) AI-powered chatbots are driven by Natural Language Processing (NLP) and Machine Learning (ML). They interpret user input in free text, extract *intent* and *entities*, and generate appropriate responses based on trained data.

Popular Tools: Google Dialogflow, IBM Watson Assistant, Microsoft LUIS.

Key Features:

- Intent detection and entity recognition
- Contextual memory across multiple turns
- Continuous learning and improvement via retraining

Drawbacks:

- Requires labeled training data
- Complex development and monitoring
- Risk of incorrect responses in sensitive domains

3. Hybrid Chatbots Hybrid chatbots combine the deterministic logic of rule-based flows with the flexibility of NLP-powered components. This makes them ideal for domains like healthcare that require both control and conversational richness.

Example in This Project:

- Structured symptom input handled via rule-based forms
- Open-ended queries processed by NLP intent classifiers
- API actions triggered through Botpress `Execute Code Card`

Benefits:

- Combines safety with user engagement
- Modular design for adaptability and scalability
- Seamless integration with predictive AI services

4. Voice-based Chatbots (Voice Assistants) Voice-based chatbots support interaction via spoken language and combine speech-to-text, NLP, and text-to-speech components.

Platforms: Amazon Alexa, Google Assistant, Apple Siri.

Use Cases:

- Elderly care support
- Medication reminders
- Hands-free symptom inquiry

Limitations:

- Requires robust speech recognition tuned for medical terms
- Noise sensitivity and privacy concerns
- Not implemented in this project but suitable for future work

Table 2: Comparison of Chatbot Types

Type	**Flexibility**	**Safety**	**NLP Support**	**Typical Use Case**
Rule-based	Low	High	No	Decision flows, FAQs
AI-powered	High	Medium	Yes	Conversational interfaces
Hybrid	Med–High	High	Yes	Healthcare triage, fintech bots
Voice-based	High	Medium	Yes	Elderly care, voice-first UI

6.4 What Platforms and Tools Can Be Used for Chatbot Integration?

The effectiveness and maintainability of a chatbot depend heavily on the chosen development platform. Table 3 summarizes leading tools:

Platform / Tool	**Description**
Botpress	Open-source, developer-friendly platform with flow editor, custom actions, and API integration. Used in this project for its flexibility and ease of web deployment.
IBM Watson Assistant	Enterprise-grade cloud chatbot solution with advanced NLU, tone analysis, and integration into healthcare systems.
Google Dialogflow	NLP-powered chatbot framework with multilingual support and integration into GCP and Google Assistant. Ideal for rapid development.
Rasa	Python-based open-source framework offering full control over the ML pipeline and deployment. Suitable for private, on-premise applications.
Microsoft Bot Framework	A suite for building bots with Azure services. Strong integration into the Microsoft ecosystem.
Messenger / Telegram / Web Chat	Channels for deployment and interaction. For this project, the `Web Chat` module is used to embed the bot into a Flask app.

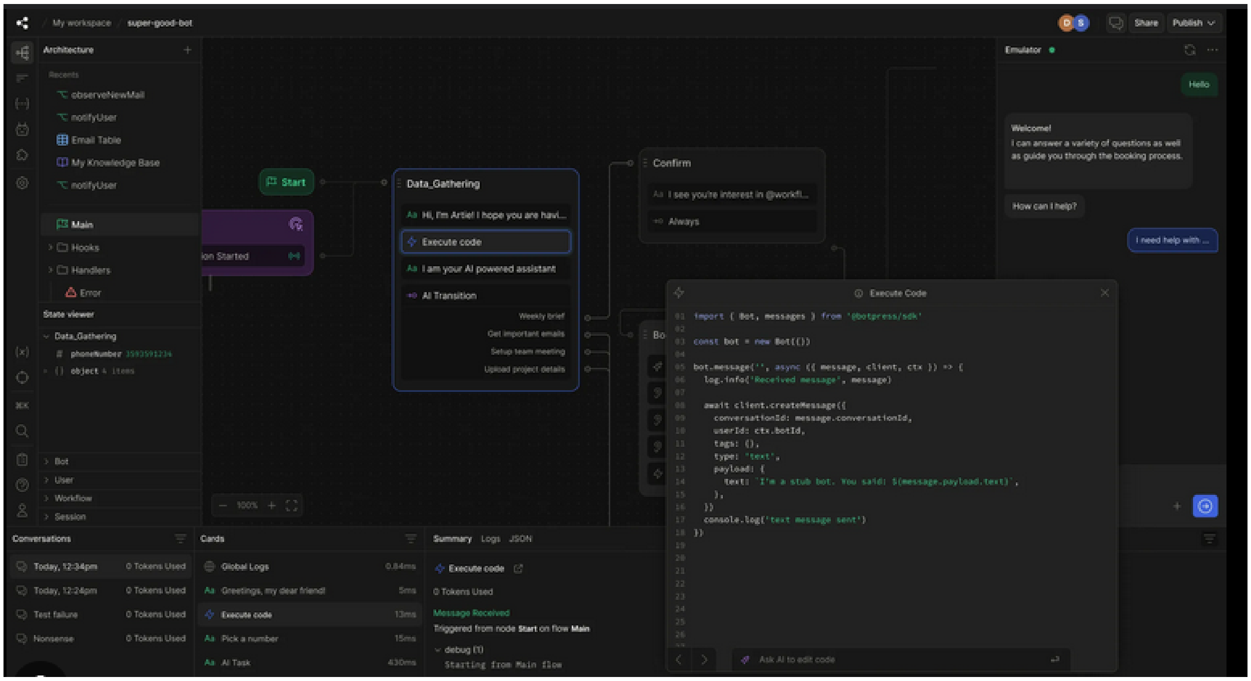
Table 3: Comparison of chatbot development platforms

6.5 Botpress: Platform Overview & Web Integration

Why Choose Botpress?

Botpress is an open-source, modular conversational AI platform well-suited for embedding AI services in structured applications such as healthcare. It combines a *visual flow editor*, *intent recognition engine*, and *custom code execution* to provide both low-code and developer-driven integration. Botpress was selected for this project due to its flexibility, ease of deployment, and support for direct API communication with a Flask backend.

Core Features



Hi, I'm Artie! I hope you are havi...

I am your AI powered assistant

I see you're interest in @workfl...

Hello

Welcome!
I can answer a variety of questions as well as guide you through the booking process.

How can I help?

I need help with ...

```
import { Bot, messages } from '@botpress/sdk'

const bot = new Bot({})

bot.message('', async ({ message, client, ctx }) => {
  log.info('Received message', message)

  await client.createMessage({
    conversationId: message.conversationId,
    userId: ctx.botId,
    tags: {},
    type: 'text',
    payload: {
      text: `I'm a stub bot. You said: ${message.payload.text}`,
    },
  })
  console.log('text message sent')
})
```

Today, 12:34pm	0 Tokens Used
Today, 12:24pm	0 Tokens Used
Test failure	0 Tokens Used
Nonsense	0 Tokens Used

Global Logs	0.84ms
Greetings, my dear friend!	5ms
Execute code	13ms
Pick a number	15ms
AI Task	430ms

```
Execute code
0 Tokens Used
Message Received
Triggered from node Start on flow Main
debug (1)
  Starting from Main flow
```

Figure 11: Botpress Visual IDE

- **Visual Flow Builder & NLU Integration**: Drag-and-drop interface to design flows, label intents, and extract entities.

- **Custom Actions**: Write JavaScript code to call APIs or perform logic during conversation.
- **Webchat SDK**: Easily embed the chatbot into a Flask-based interface.
- **Multi-channel Deployment**: Facebook, Telegram, or Web.

Web Integration with Botpress

To embed the Botpress Webchat in a page (e.g., `base.html`):

Listing 10: Botpress Webchat HTML Snippet

```
<!DOCTYPE html>
<html>
<head>
  <script src="http://<BOTPRESS_HOST>:3000/assets/
     modules/channel-web/inject.js"></script>
</head>
<body>
  <script>
    window.botpressWebChat.init({
      host: ’http://<BOTPRESS_HOST>:3000’,
      botId: ’<YOUR_BOT_ID>’,
      containerWidth: ’350px’,
      layoutWidth: ’350px’,
      showPoweredBy: false
    })
  </script>
  {% block content %}{% endblock %}
</body>
</html>
```

Custom Botpress Action

Define a custom action in Botpress Studio :

Listing 11: Botpress Custom Action

```
const axios = require(’axios’);

async function callPredictRisk(state, event, { apiUrl })
    {
  const payload = {
    age: state.age,
    smoking: state.smoking,
    shortness_of_breath: state.shortness_of_breath
  };

  try {
    const resp = await axios.post(apiUrl, payload);
    state.risk_level = resp.data.risk_level;
    return state;
  } catch (err) {
    event.reply(’#builtin_text’, { text: ’Error
       computing risk’ });
    return state;
  }
}

return callPredictRisk;
```

7 Application Development

7.1 Overview

The lung cancer prediction system was deployed as a web application using the **Streamlit** framework. This app provides a comprehensive interface for user interaction, combining:

- Advanced **machine learning models** for cancer risk prediction,
- **Exploratory Data Analysis (EDA)** for visual insights, and
- A natural language-based **chatbot** for educational support.

This integration delivers a dual function: early-stage diagnostic support and public health education.

Project Structure

```
    app       .py                          # Main Streamlit
  app controller
    chatbot       .py                      # Chatbot
interface and response logic
    eda       .py                          # EDA
visualizations and analysis
     introduction        .py               # Introductory
content from Markdown files
     prediction        .py                 # Prediction
logic and form interface
     visualization        .py              # Plotting
utility functions
    info       /                           # Static content
          datades       .md                  # Dataset
description
          intro       .md                    # Project
introduction
    model       /                          # Pre-trained ML
models
```

This modular structure ensures clear separation of concerns between UI rendering, model prediction, data visualization, and chatbot communication.

7.2 Core Functionalities

The application is organized into four main tabs, each targeting a specific function:

- **Introduction**: Presents an overview of the project goals and dataset characteristics.

- **EDA (Exploratory Data Analysis)**: Allows users to explore patterns and insights within the lung cancer dataset through interactive visualizations.
- **Predict**: Enables users to input their health parameters and receive a lung cancer risk prediction using trained ML models.
- **Chatbot**: Provides a conversational agent that answers questions related to lung cancer using AI-driven natural language processing.

7.3 Technical Architecture

The system architecture follows a modular design:

- `app.py`: Main driver of the Streamlit app, managing tab navigation and orchestrating individual module executions.
- `prediction.py`: Handles the predictive interface. It includes a user form, loads the pre-trained Random Forest and AdaBoost models, and visualizes risk results.
- `eda.py`: Implements EDA visualizations, calling reusable plotting functions from `visualization.py`.
- `introduction.py`: Displays static markdown content introducing the project, loaded from files in the `info/` directory.
- `chatbot.py`: Manages the chatbot's logic, including interface design and real-time response generation using a language model.
- `visualization.py`: Contains functions for rendering charts like bar plots, histograms, and correlation heatmaps.
- `model/`: Stores serialized machine learning models in `.pkl` format.
- `info/`: Stores Markdown documents (`intro.md`, `datades.md`) displayed in the app.

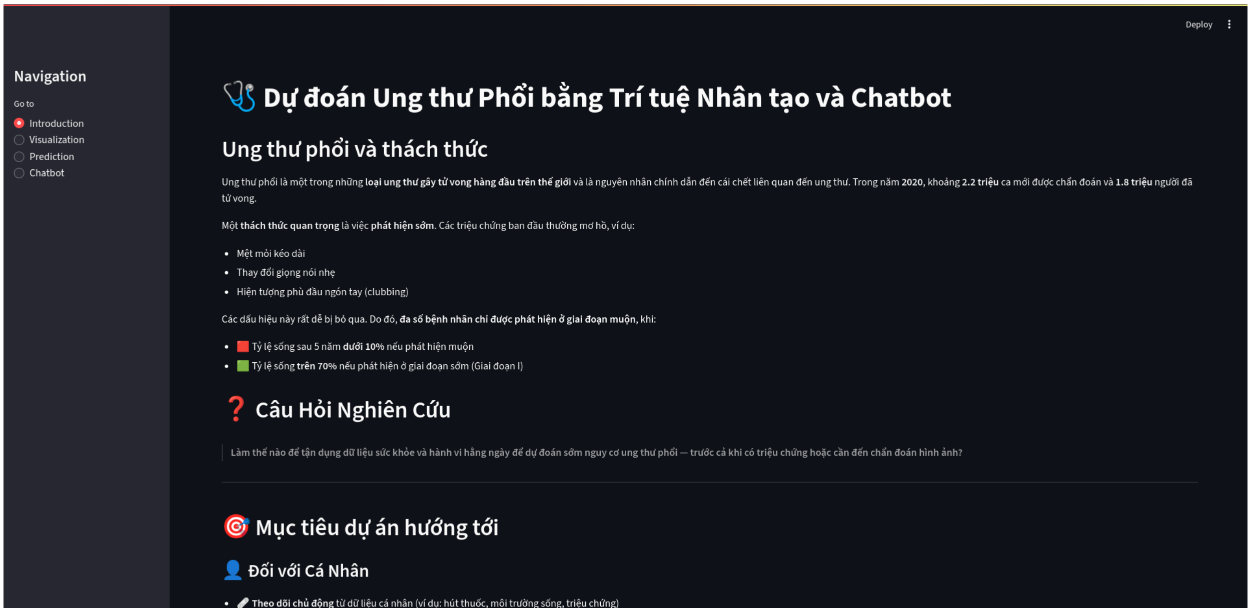
This architecture promotes separation of concerns, enhances maintainability, and simplifies future scalability.

7.4 User Interface and Experience

Built with Streamlit's multi-page layout and `streamlit_option_menu`, the interface is intuitive and responsive. Key UI considerations include:

- Clear tab navigation with labeled sections.
- Simple forms for inputting prediction parameters.
- Real-time interaction with the AI chatbot.
- Visual storytelling via dynamic charts and plots.

7.4.1 Introduction Tab



🩺 Dự đoán Ung thư Phổi bằng Trí tuệ Nhân tạo và Chatbot

Ung thư phổi và thách thức

Ung thư phổi là một trong những **loại ung thư gây tử vong hàng đầu trên thế giới** và là nguyên nhân chính dẫn đến cái chết liên quan đến ung thư. Trong năm **2020**, khoảng **2.2 triệu** ca mới được chẩn đoán và **1.8 triệu** người đã tử vong.

Một **thách thức quan trọng** là việc **phát hiện sớm**. Các triệu chứng ban đầu thường mơ hồ, ví dụ:

- Mệt mỏi kéo dài
- Thay đổi giọng nói nhẹ
- Hiện tượng phù đầu ngón tay (clubbing)

Các dấu hiệu này rất dễ bị bỏ qua. Do đó, **đa số bệnh nhân chỉ được phát hiện ở giai đoạn muộn**, khi:

- 🟥 Tỷ lệ sống sau 5 năm **dưới 10%** nếu phát hiện muộn
- 🟩 Tỷ lệ sống **trên 70%** nếu phát hiện ở giai đoạn sớm (Giai đoạn I)

❓ Câu Hỏi Nghiên Cứu

> Làm thế nào để tận dụng dữ liệu sức khỏe và hành vi hằng ngày để dự đoán sớm nguy cơ ung thư phổi — trước cả khi có triệu chứng hoặc cần đến chẩn đoán hình ảnh?

🎯 Mục tiêu dự án hướng tới

👤 Đối với Cá Nhân

- 🩹 **Theo dõi chủ động** từ dữ liệu cá nhân (ví dụ: hút thuốc, môi trường sống, triệu chứng)

Figure 12: Introduction Tab Interface of the Application

The **Introduction** tab serves as the landing page of the web application. It aims to communicate the background, motivation, and objectives of the lung cancer prediction project in a concise yet informative manner. This section plays an important role in orienting users, including both medical professionals and laypersons, toward the problem being addressed and the solution being proposed.

Interface Design and Implementation The introduction page is rendered using the `show_introduction()` function defined in `introduction.py`. This function loads and displays content from a static Markdown file (`intro.md`) located in the `info/` directory. Streamlit's `st.markdown` function is used to format and render the content, allowing for structured text with headers, bullet points, and emphasis.

Content Overview The introductory content is structured into the following thematic sections:

1. **Context and Problem Statement**
   This section introduces lung cancer as one of the most fatal forms of cancer globally. It references 2020 statistics, where approximately 2.2 million new cases were diagnosed and 1.8 million deaths were reported. The content emphasizes the critical challenge of early detection, highlighting that symptoms in the initial stages—such as fatigue, minor voice changes, or subtle physical signs—are often overlooked. Consequently, a large proportion of cases are only diagnosed at late stages, drastically reducing survival rates.

2. **Survival Rate Comparison**
   The page contrasts survival outcomes between early and late detection. Specifically, it notes that five-year survival rates drop below 10% when cancer is detected late, compared to rates above 70% when detected in Stage I. This contrast underscores the importance of early intervention and the need for accessible, scalable screening tools.

3. **Research Question**
   The introduction raises the central research question guiding the project: how can personal health and behavioral data be leveraged to predict lung cancer risk before traditional symptoms or imaging-based diagnostics are required?

4. **Project Objectives and Stakeholder Impact**
   The tab outlines specific benefits tailored to three stakeholder groups:

   - **Individuals**: Encouraging proactive monitoring through personal health data and promoting behavioral change prior to clinical diagnosis.
   - **Healthcare Providers**: Enabling data-driven risk stratification and prioritization for advanced screening methods like low-dose CT.
   - **Healthcare Systems**: Reducing the diagnostic burden through AI-driven preliminary assessment and shifting resources toward prevention.

5. **Long-Term Vision**
   The final section of the introduction articulates the broader ambitions of the project, including the advancement of digital health innovation and patient empowerment. It emphasizes the integration of machine learning, chatbot technology, and accessible web platforms to offer a cost-effective, scalable, and user-centric health solution.

7.4.2 EDA Tab

The Exploratory Data Analysis (EDA) tab provides users with interactive tools to explore and understand the underlying lung cancer dataset used in model training and analysis. It is designed to enhance transparency, promote user engagement with the data, and support interpretability of the system's predictive components.

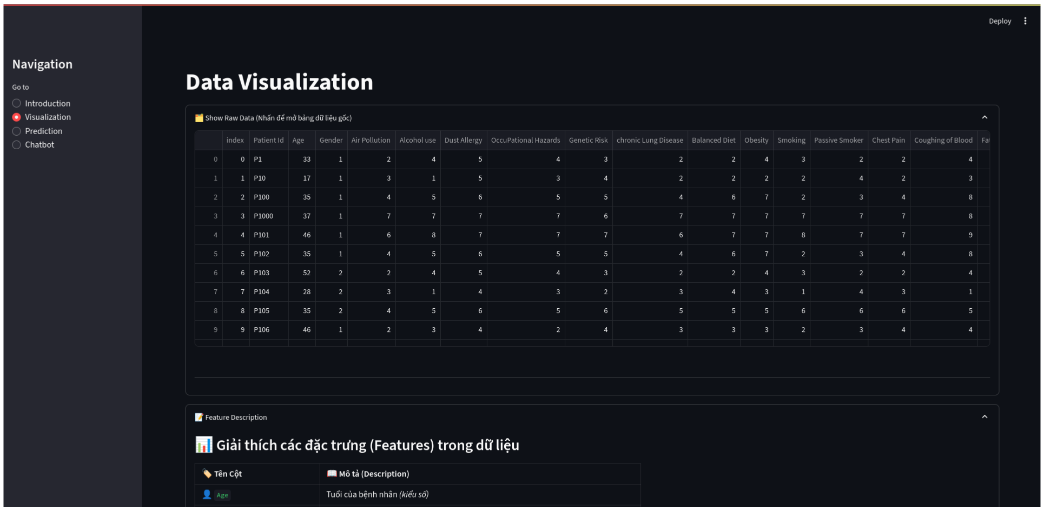
Functional Overview This tab consists of four primary interactive components, each serving a distinct data exploration purpose:

1. Raw Data Viewer
2. Feature Description Panel
3. Feature Comparison Module

4. Correlation Heatmap

Each component is accessible via collapsible sections and is implemented using Streamlit's UI elements and Python plotting libraries such as `matplotlib` and `seaborn`.

1. Raw Data Viewer This section displays the full dataset used in model development, allowing users to inspect individual records and understand the structure and values associated with each patient. The data table includes attributes such as patient ID, age, gender, environmental exposure levels, lifestyle factors, and clinical symptoms.



Data Visualization

🗂️ Show Raw Data (Nhấn để mở bảng dữ liệu gốc)

	index	Patient Id	Age	Gender	Air Pollution	Alcohol use	Dust Allergy	OccuPational Hazards	Genetic Risk	chronic Lung Disease	Balanced Diet	Obesity	Smoking	Passive Smoker	Chest Pain	Coughing of Blood	Fa
0	0	P1	33	1	2	4	5	4	3	2	2	4	3	2	2	4	
1	1	P10	17	1	3	1	5	3	4	2	2	2	2	4	2	3	
2	2	P100	35	1	4	5	6	5	5	4	6	7	2	3	4	8	
3	3	P1000	37	1	7	7	7	7	6	7	7	7	7	7	7	8	
4	4	P101	46	1	6	8	7	7	7	6	7	7	8	7	7	9	
5	5	P102	35	1	4	5	6	5	5	4	6	7	2	3	4	8	
6	6	P103	52	2	2	4	5	4	3	2	2	4	3	2	2	4	
7	7	P104	28	2	3	1	4	3	2	3	4	3	1	4	3	1	
8	8	P105	35	2	4	5	6	5	6	5	5	5	6	6	6	5	
9	9	P106	46	1	2	3	4	2	4	3	3	3	2	3	4	4	

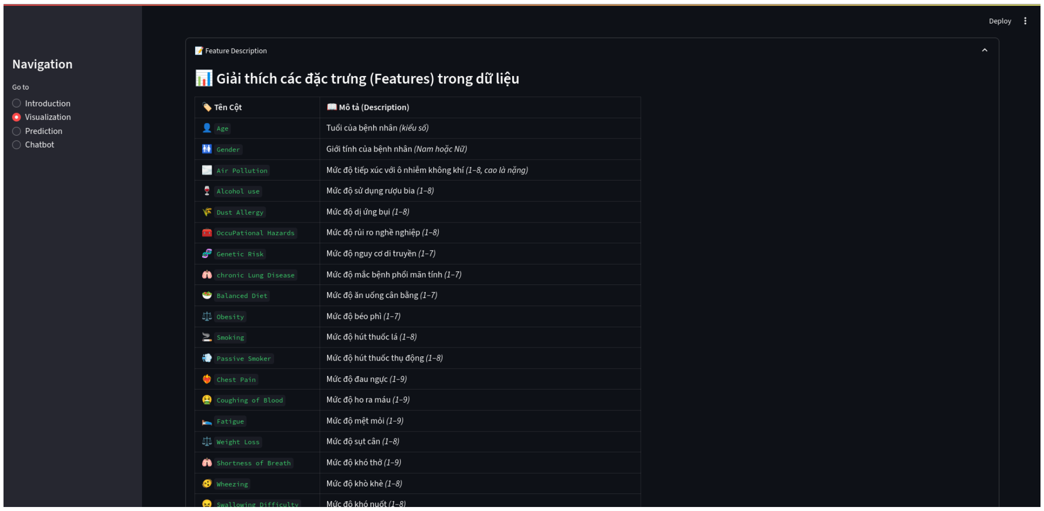
📝 Feature Description

📊 Giải thích các đặc trưng (Features) trong dữ liệu

🏷️ Tên Cột	📖 Mô tả (Description)
👤 Age	Tuổi của bệnh nhân *(kiểu số)*

Figure 13: Raw data preview interface with scrollable table

2. Feature Description Panel To support interpretability, this section provides a detailed description of each feature in the dataset. The panel is rendered as a two-column table: one listing the feature names, and the other offering corresponding definitions, value ranges, and units where applicable.



📝 Feature Description

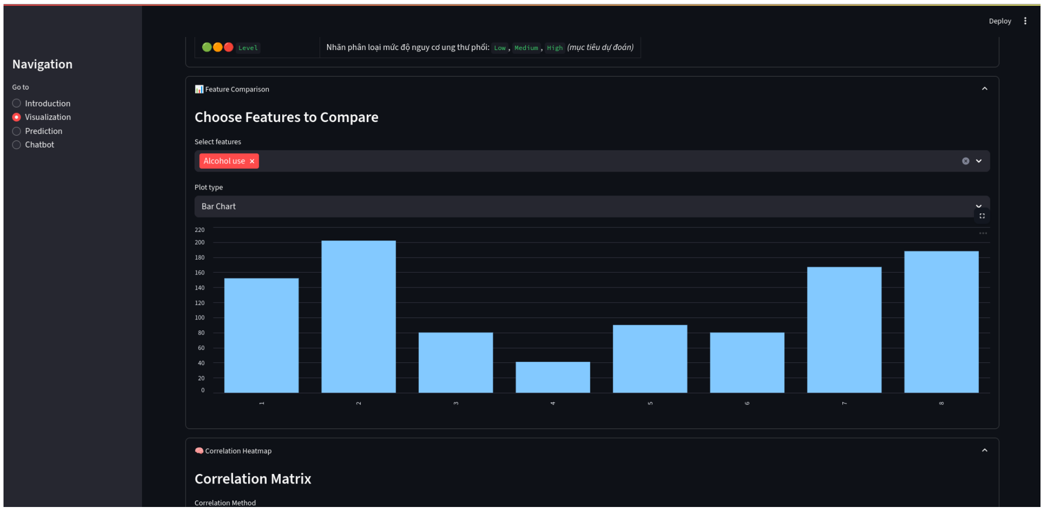
📊 Giải thích các đặc trưng (Features) trong dữ liệu

🏷️ Tên Cột	📖 Mô tả (Description)
👤 `Age`	Tuổi của bệnh nhân *(kiểu số)*
🚻 `Gender`	Giới tính của bệnh nhân *(Nam hoặc Nữ)*
🌫️ `Air Pollution`	Mức độ tiếp xúc với ô nhiễm không khí *(1–8, cao là nặng)*
🍷 `Alcohol use`	Mức độ sử dụng rượu bia *(1–8)*
🌾 `Dust Allergy`	Mức độ dị ứng bụi *(1–8)*
🧰 `OccuPational Hazards`	Mức độ rủi ro nghề nghiệp *(1–8)*
🧬 `Genetic Risk`	Mức độ nguy cơ di truyền *(1–7)*
🫁 `chronic Lung Disease`	Mức độ mắc bệnh phổi mãn tính *(1–7)*
🥗 `Balanced Diet`	Mức độ ăn uống cân bằng *(1–7)*
⚖️ `Obesity`	Mức độ béo phì *(1–7)*
🚬 `Smoking`	Mức độ hút thuốc lá *(1–8)*
💨 `Passive Smoker`	Mức độ hút thuốc thụ động *(1–8)*
❤️‍🔥 `Chest Pain`	Mức độ đau ngực *(1–9)*
🤮 `Coughing of Blood`	Mức độ ho ra máu *(1–9)*
🛌 `Fatigue`	Mức độ mệt mỏi *(1–9)*
⚖️ `Weight Loss`	Mức độ sụt cân *(1–8)*
🫁 `Shortness of Breath`	Mức độ khó thở *(1–9)*
😮‍💨 `Wheezing`	Mức độ khò khè *(1–8)*
😖 `Swallowing Difficulty`	Mức độ khó nuốt *(1–8)*

Figure 14: Feature dictionary explaining each input variable

3. Feature Comparison Module This section enables users to select one or more features and visualize their distribution across the dataset using various chart types such as bar plots and histograms. The purpose is to examine how certain characteristics vary among patients and potentially relate to cancer risk.

Users can dynamically specify the feature of interest and the visualization style. For instance, Figure 15 illustrates the distribution of alcohol use scores among patients.



🟢🟠🔴 `Level` Nhãn phân loại mức độ nguy cơ ung thư phổi: `Low`, `Medium`, `High` *(mục tiêu dự đoán)*

Choose Features to Compare

Select features: Alcohol use ×

Plot type: Bar Chart

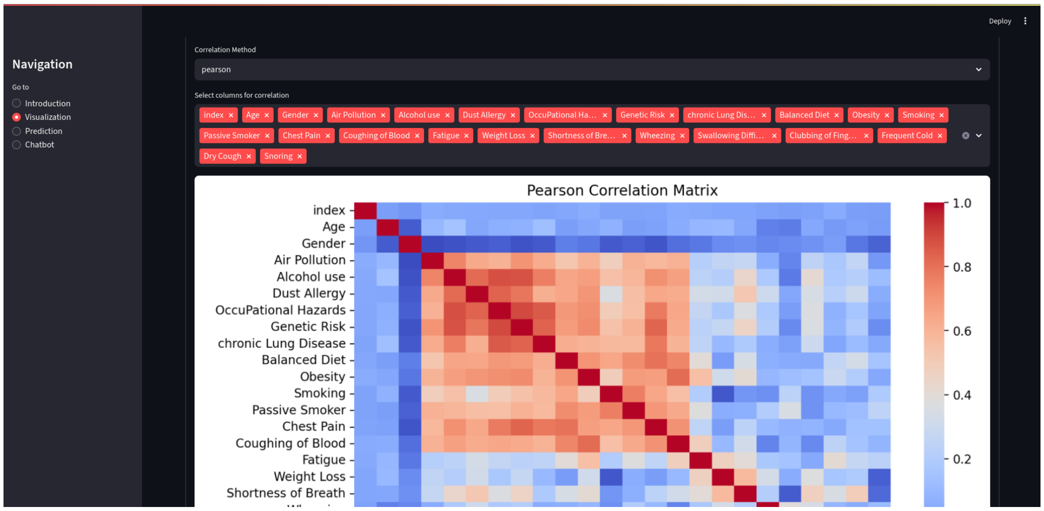
Correlation Matrix

Correlation Method

Figure 15: Feature comparison using bar charts

4. Correlation Heatmap This module presents a visual correlation matrix that allows users to examine linear relationships between multiple features. It supports correlation computation via different statistical methods (e.g., Pearson, Spearman) and provides visual cues via a color gradient to indicate correlation strength and direction.

As shown in Figure 16, users can select a subset of features for comparison. The resulting matrix highlights pairs of variables with strong positive or negative correlations, which can inform feature selection or interpretation in downstream modeling.



Go to

○ Introduction
● Visualization
○ Prediction
○ Chatbot

Correlation Method: pearson

Select columns for correlation: index × Age × Gender × Air Pollution × Alcohol use × Dust Allergy × OccuPational Ha… × Genetic Risk × chronic Lung Dis… × Balanced Diet × Obesity × Smoking × Passive Smoker × Chest Pain × Coughing of Blood × Fatigue × Weight Loss × Shortness of Bre… × Wheezing × Swallowing Diffi… × Clubbing of Fing… × Frequent Cold × Dry Cough × Snoring ×

Figure 16: Pearson correlation matrix for selected features

The EDA tab offers a comprehensive suite of visualization tools that make data inspection intuitive and informative. It plays a critical role in supporting data transparency, educating users on dataset composition, and revealing relationships that may inform predictive model behavior. Each subcomponent contributes to a more explainable and trustworthy AI system for lung cancer risk assessment.

7.4.3 Prediction Tab

The **Prediction** tab is the central component of the application, offering users the ability to obtain real-time lung cancer risk predictions based on individual health data. It combines model inference with human-centered explainability to deliver both quantitative outputs and qualitative insights.

Functional Architecture The prediction interface is built using the `show_prediction()` function in `prediction.py`. It integrates model loading, dynamic form generation, user input capture, prediction display, and explainability via SHAP visualizations.

This tab includes the following interactive components:

1. Model Selection

2. Feature Importance Display

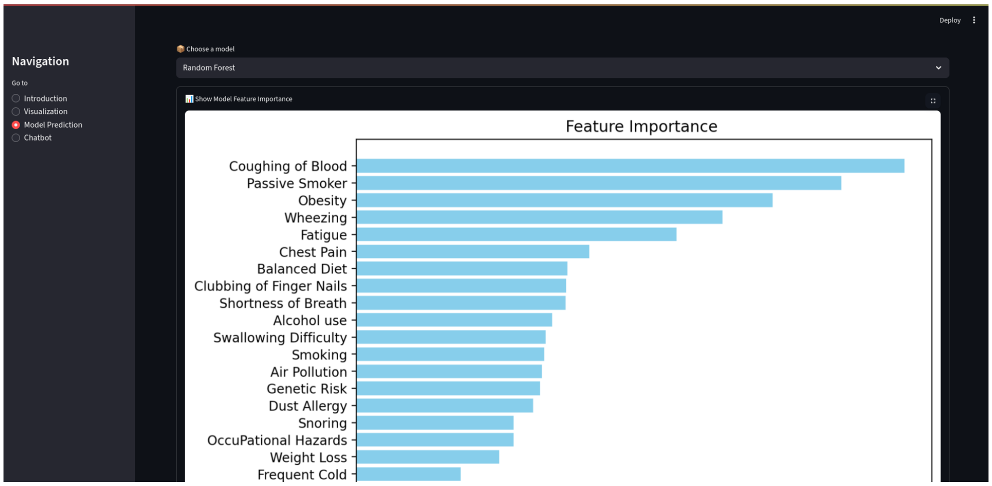
3. User Input Form

4. Prediction Result Output

5. SHAP-Based Explainability

1. Model Selection Users can choose between two pre-trained machine learning models. Once a model is selected, the corresponding `.pkl` file is dynamically loaded using `joblib`. This approach supports comparative testing and provides flexibility for experimenting with different classifiers.

2. Feature Importance Display For models that support it (e.g., Random Forest), a global feature importance plot is provided. This horizontal bar chart visualizes the relative contribution of each input feature to the overall model prediction.



📦 Choose a model: Random Forest

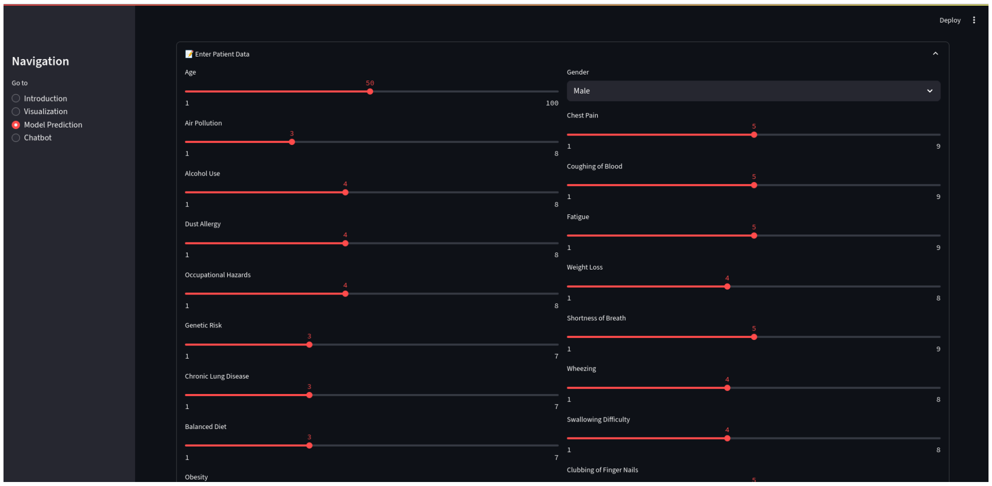
Figure 17: Feature importance visualization for Random Forest model

If the model does not support feature importance extraction, the application will suppress this section and display an informative message.

3. User Input Form The core of the prediction tab is an interactive input form that collects detailed information across 23 health and demographic attributes, including:

- **Demographic**: Age, Gender
- **Environmental**: Air Pollution, Occupational Hazards
- **Lifestyle**: Smoking, Alcohol use, Passive Smoker, Balanced Diet
- **Medical History and Symptoms**: Chest Pain, Chronic Lung Disease, Genetic Risk, Fatigue, etc.

Each field is implemented using sliders or dropdowns to ensure structured input. The layout uses two-column vertical alignment for readability and ease of interaction.



📝 Enter Patient Data

Age
50
1 100

Air Pollution
3
1 8

Alcohol Use
4
1 8

Dust Allergy
4
1 8

Occupational Hazards
4
1 8

Genetic Risk
3
1 7

Chronic Lung Disease
3
1 7

Balanced Diet
3
1 7

Obesity

Gender
Male

Chest Pain
5
1 9

Coughing of Blood
5
1 9

Fatigue
5
1 9

Weight Loss
4
1 8

Shortness of Breath
5
1 9

Wheezing
4
1 8

Swallowing Difficulty
4
1 8

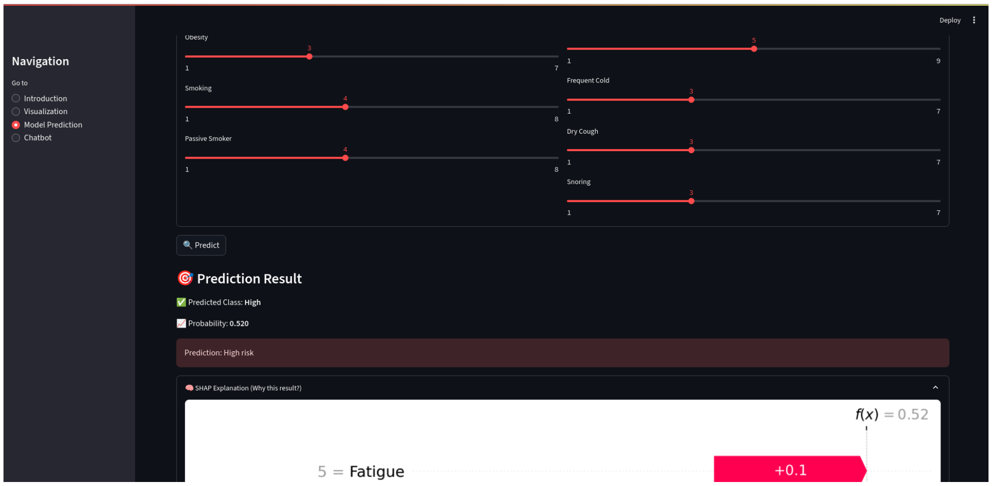
Clubbing of Finger Nails
5

Figure 18: User input interface with structured feature entry

4. Prediction Result Output Upon clicking the **Predict** button, the application executes the following steps:

- Computes the predicted class (Low, Medium, or High risk)
- Retrieves the corresponding class probability

- Displays the result with color-coded emphasis:
  - Green for Low risk
  - Yellow for Medium risk
  - Red for High risk



Obesity 3 (1–7)

Smoking 4 (1–8)

Passive Smoker 4 (1–8)

5 (1–9)

Frequent Cold 3 (1–7)

Dry Cough 3 (1–7)

Snoring 3 (1–7)

🎯 Prediction Result

✅ Predicted Class: **High**

📈 Probability: **0.520**

Prediction: High risk

🧠 SHAP Explanation (Why this result?)

Figure 19: Prediction result with probability and risk category

This visual feedback mechanism ensures interpretability for both clinical and general users.

5. SHAP-Based Explainability To enhance interpretability at the individual level, the system integrates **SHAP (SHapley Additive exPlanations)**. Key components include:

- Use of `TreeExplainer` for decision tree-based models
- Generation of SHAP waterfall plots showing feature contributions
- Expandable display interface for explanation results

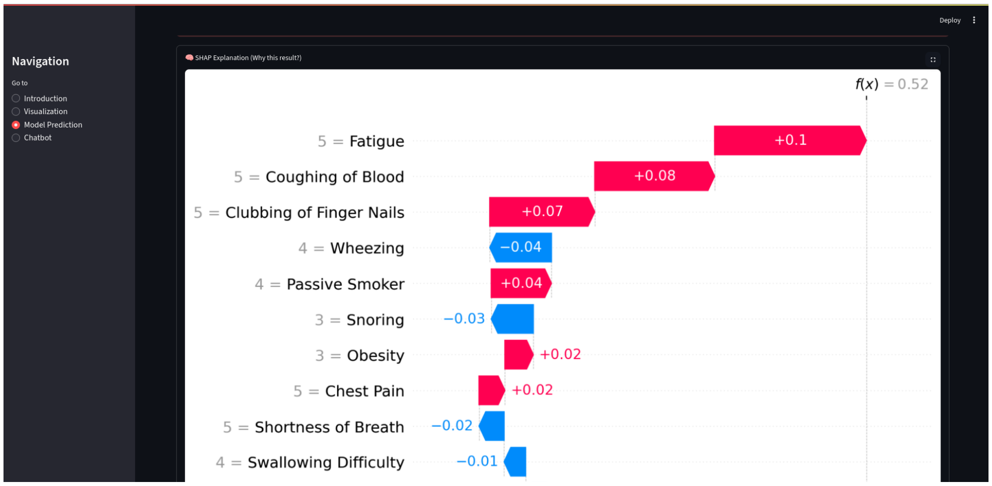


Figure 20: SHAP explanation plot showing feature contributions for individual prediction

This explainability component is essential for fostering trust and providing clinical justification for AI-driven decisions.

The Prediction tab transforms the underlying machine learning pipeline into a user-accessible, explainable diagnostic tool. By combining structured input, flexible model choice, immediate result feedback, and SHAP-based interpretation, this tab provides a robust and transparent AI-enabled risk assessment interface. It plays a pivotal role in operationalizing AI for early lung cancer screening and education.

7.4.4 Chatbot Tab

The **Chatbot** tab integrates an AI-powered conversational agent designed to assist users in understanding key aspects of lung cancer. It functions as an accessible, real-time educational tool to answer health-related questions, clarify predictions, and provide guidance on prevention and risk factors.

System Integration The chatbot is implemented using the `show_chatbot()` function defined in `chatbot.py`. The integration is achieved by embedding a web-based Botpress agent within the Streamlit application through an

`<iframe>` HTML component. This allows for seamless interaction without redirecting users to external sites.

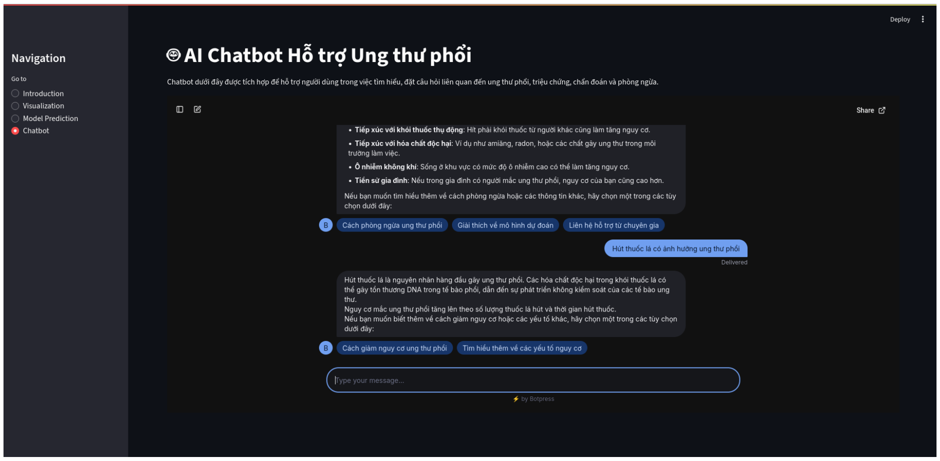
The chatbot configuration is hosted remotely and fetched dynamically using a public `configUrl` from Botpress Cloud. This architecture ensures maintainability and allows content and logic updates without altering the frontend code.

Functional Capabilities The chatbot provides multi-purpose conversational support in the following areas:

- Lung cancer prevention strategies
- Symptom explanations and risk factor education
- Interpretation of model predictions
- Referrals to expert consultation and screening recommendations

It supports both **free-form question answering** and **button-based suggestions** to guide users through complex queries.

Interface and User Experience The Chatbot tab is designed with simplicity and responsiveness in mind. Upon accessing the tab, users are greeted with a title and brief instruction on how to use the assistant. The chatbot is presented in a large embedded panel occupying the majority of the page width.


Deploy
Navigation
Go to
○ Introduction
○ Visualization
○ Model Prediction
● Chatbot
Share
B
Cách phòng ngừa ung thư phổi
Giải thích về mô hình dự đoán
Liên hệ hỗ trợ từ chuyên gia
Delivered
B
Cách giảm nguy cơ ung thư phổi
Tìm hiểu thêm về các yếu tố nguy cơ
Type your message...
⚡ by Botpress

AI Chatbot Hỗ trợ Ung thư phổi

Chatbot dưới đây được tích hợp để hỗ trợ người dùng trong việc tìm hiểu, đặt câu hỏi liên quan đến ung thư phổi, triệu chứng, chẩn đoán và phòng ngừa.

- **Tiếp xúc với khói thuốc thụ động**: Hít phải khói thuốc từ người khác cũng làm tăng nguy cơ.
- **Tiếp xúc với hóa chất độc hại**: Ví dụ như amiăng, radon, hoặc các chất gây ung thư trong môi trường làm việc.
- **Ô nhiễm không khí**: Sống ở khu vực có mức độ ô nhiễm cao có thể làm tăng nguy cơ.
- **Tiền sử gia đình**: Nếu trong gia đình có người mắc ung thư phổi, nguy cơ của bạn cũng cao hơn.

Nếu bạn muốn tìm hiểu thêm về cách phòng ngừa hoặc các thông tin khác, hãy chọn một trong các tùy chọn dưới đây:

Hút thuốc lá có ảnh hưởng ung thư phổi

Hút thuốc lá là nguyên nhân hàng đầu gây ung thư phổi. Các hóa chất độc hại trong khói thuốc lá có thể gây tổn thương DNA trong tế bào phổi, dẫn đến sự phát triển không kiểm soát của các tế bào ung thư.
Nguy cơ mắc ung thư phổi tăng lên theo số lượng thuốc lá hút và thời gian hút thuốc.
Nếu bạn muốn biết thêm về cách giảm nguy cơ hoặc các yếu tố khác, hãy chọn một trong các tùy chọn dưới đây:

Figure 21: Embedded chatbot assistant interface with contextual guidance.

Users can enter questions via a message box or select from contextual buttons related to common topics such as:

- Lung cancer risk factors
- Prevention methods
- Symptom understanding
- Model explanation

Responses are delivered in a structured text format, providing medically relevant but accessible answers.

Role in the Application This module plays an essential role in making the system not only a diagnostic tool but also a **health education platform**. Unlike static dashboards, the chatbot enables natural language interaction, helping users find answers to their unique concerns and encouraging proactive engagement in their health journey.

Moreover, by integrating AI explanations and dynamic health advice, the chatbot helps address concerns of transparency, user empowerment, and accessibility in healthcare applications.

7.5 Installation and Deployment

The application can be set up with the following steps:

1. Clone the repository and navigate to the source directory:

```
git clone <repository-url>
cd lung_cancer_project/src/application
```

2. Set up a virtual environment:

```
python -m venv venv
source venv/bin/activate  # Windows: venv\Scripts\
   activate
```

3. Install required packages:

```
pip install -r requirements.txt
```

4. Launch the app:

```
streamlit run app.py
```

8 Conclusion

This project presents a comprehensive, end-to-end implementation of an AI-powered system for lung cancer risk prediction, built upon rigorous data analysis, machine learning modeling, and interactive application development. By leveraging a rich dataset comprising demographic, behavioral, environmental, genetic, and clinical attributes of 1,000 patients, we successfully identified key risk factors that contribute to lung cancer susceptibility—most notably smoking behavior, chronic lung disease, and genetic predisposition.

A series of Exploratory Data Analysis (EDA) visualizations revealed distinct patterns in risk factor distributions across disease severity levels. Notably, features such as **Genetic Risk** and **Chronic Lung Disease** demonstrated strong discriminative power, especially among High-risk patients, and exhibited consistent monotonic trends. These insights provided a solid foundation for feature selection and model interpretation.

Two tree-based ensemble methods—**Random Forest Classifier** and **Gradient Boosting Classifier**—were applied to classify patients into Low, Medium, and High lung cancer risk categories. These models were chosen for their ability to capture nonlinear interactions and handle heterogeneous feature distributions. Performance was evaluated using a suite of classification metrics, including accuracy, precision, recall, F1-score, confusion matrix, and ROC-AUC. The results (detailed in Section **??**) demonstrate the model's strong potential for reliable and explainable risk prediction in a healthcare context.

Beyond modeling, the system was extended with a **chatbot component**, implemented using Botpress, to facilitate natural language interaction and user engagement. This conversational agent not only collects symptom and lifestyle data from users but also delivers risk predictions and health guidance in real-time—effectively democratizing access to AI-driven insights. By embedding the chatbot into a **Streamlit web interface**, the system becomes highly accessible, scalable, and integrable into broader digital health ecosystems.

Finally, a web application was developed, integrating prediction services, visualization dashboards, feature interpretability tools, and the chatbot interface. The application serves as a robust and extensible decision-support platform—one that empowers users, supports clinicians, and promotes early intervention through personalized risk assessment.

Future Work

While the current system demonstrates strong potential, several avenues remain open for future development:

- **Expanding the dataset** with longitudinal or external patient records for better validation and generalizability.
- **Enhancing model robustness** through hyperparameter optimization, ensemble stacking, or deep learning architectures.

- **Integrating medical imaging data** (e.g., chest CT scans) to enable multimodal risk prediction.
- **Voice-enabled chatbot extension** to improve accessibility, particularly for elderly users or those with visual impairments.
- **Cloud deployment** with user authentication, session management, and end-to-end data security protocols for real-world usage.

Final Remarks

In sum, this project exemplifies the integration of data science, AI modeling, conversational agents, and software engineering in the service of predictive, preventive, and participatory healthcare. By building a scalable and interpretable system for lung cancer risk assessment, we contribute a concrete step toward early detection and health equity through intelligent, user-centric technology.