Lung Cancer Survival Prediction System: A Machine Learning Approach

Author: Parveen Kashyap
Internship ID: UMID13062542814
Organization: Unified Mentor Pvt. Ltd.
Duration: 6 months (15 June 2025 – 15 December 2025)

Abstract

This project presents a machine learning system designed to predict the survival likelihood of lung cancer patients using patient medical history, diagnosis details, and treatment information. Utilizing a Random Forest Classifier with 200 estimators, the system analyzes 14 key clinical features to provide survival predictions. The model achieved an accuracy of approximately 90% through comprehensive data preprocessing, feature engineering, and stratified validation on an 80-20 train-test split. The system integrates an interactive web interface built with Streamlit to enable real-time predictions and risk factor visualization. This work demonstrates the practical application of machine learning in medical prognosis support and highlights the significance of clinical factors in patient outcome prediction.

Introduction

Lung cancer remains one of the most prevalent and deadly malignancies worldwide, necessitating accurate prognostic tools to support clinical decision-making. Early identification of high-risk patients can significantly impact treatment planning and resource allocation. Traditional prognostic methods rely on clinician expertise and historical patient data; however, machine learning approaches can systematically analyze complex relationships across multiple clinical variables to improve prediction accuracy[1].

This project develops a predictive system leveraging patient medical data to forecast survival outcomes. By combining structured patient information with advanced machine learning techniques, the system aims to provide clinicians with data-driven insights to support patient assessment and treatment planning. The implementation includes both a trained model and an interactive visualization platform for end-user accessibility.

Problem Statement

Healthcare professionals require objective, data-driven tools to assess patient survival risk based on clinical parameters. Current diagnostic approaches often depend on individual expertise and experience, which can be inconsistent and time-consuming. The challenge

addressed in this project is to build a reproducible, automated system that accurately predicts lung cancer patient survival likelihood based on readily available clinical and demographic variables.

Specifically, the system must:

- Accurately classify patients into survival categories (survived vs. did not survive)
- Identify key clinical factors most influential in survival outcomes
- Provide confidence scores to quantify prediction reliability
- Present results through an intuitive interface for non-technical users

Dataset Description

The dataset comprises comprehensive patient information from individuals diagnosed with lung cancer. The dataset contains 16 clinical and demographic variables documented at diagnosis and treatment conclusion.

Feature	Description
age	Patient age at diagnosis (years)
gender	Patient gender (Male/Female)
country	Country of patient residence
cancer_stage	Severity stage (I, II, III, IV)
family_history	Family history of cancer (Yes/No)
smoking_status	Smoking status (Current, Former, Never, Passive)
bmi	Body Mass Index at diagnosis
cholesterol_level	Serum cholesterol level (mg/dL)
hypertension	Presence of high blood pressure (Yes/No)
asthma	History of asthma condition (Yes/No)
cirrhosis	History of liver cirrhosis (Yes/No)
other_cancer	History of other cancer types (Yes/No)
treatment_type	Treatment modality (Surgery, Chemotherapy, Radiation, Combined)
treatment_duration	Duration of treatment in days (engineered feature)
survived	Target variable (0 = did not survive, 1 = survived)

Table 1: Dataset Features and Descriptions

The dataset captures both demographic factors (age, gender, country) and clinical variables (stage, comorbidities, treatment type) that influence survival outcomes. Treatment duration

was engineered from diagnosis and treatment end dates to capture temporal treatment factors.

Methodology

Data Preprocessing

Data preparation involved multiple sequential steps to ensure model readiness:

Feature Engineering: Treatment duration (in days) was calculated from diagnosis date and treatment end date to quantify treatment continuity and intensity.

Cleaning: Non-predictive columns (patient ID) and raw date columns were removed after feature extraction.

Categorical Encoding:

- Ordinal categorical variables (cancer stage) and nominal variables (gender, country, smoking status) were encoded using LabelEncoder
- Binary conditions (hypertension, asthma, cirrhosis, other cancer, family history) were converted to numeric format (0/1)

Imputation: Missing values were handled using median imputation for numeric features and mode imputation for categorical features.

Normalization: StandardScaler was applied to normalize all features to zero mean and unit variance, ensuring consistent feature contribution regardless of original scale.

Model Architecture

- **Algorithm**: Random Forest Classifier
- **Number of Estimators**: 200 decision trees
- **Maximum Depth**: 15 levels per tree
- **Minimum Samples Split**: 5 samples required to split a node
- **Validation Strategy**: Stratified 80-20 train-test split preserving class distribution

Random Forest was selected for its robustness to non-linear relationships, ability to handle mixed data types, resistance to overfitting through ensemble averaging, and inherent feature importance calculation capabilities. The model processes each feature through 200 independent decision trees and aggregates predictions through majority voting.

Model Explanation

Random Forest Classifier

A Random Forest model consists of multiple decision trees trained on random subsets of data and features. Each tree independently learns to classify patients, and final predictions result from aggregating individual tree outputs through majority voting for classification tasks.

Advantages of this approach:

1. Handles non-linear feature relationships without explicit specification
2. Provides inherent feature importance rankings
3. Robust to outliers and missing data handling
4. Reduces overfitting through ensemble averaging across diverse trees
5. Maintains interpretability through decision tree logic

Model Parameters Explained:

- **n_estimators (200)**: Higher tree count increases model stability and prediction consistency
- **max_depth (15)**: Limits tree complexity to prevent overfitting while maintaining expressiveness
- **min_samples_split (5)**: Prevents trees from creating nodes based on minimal data points

The trained model learns which patient characteristics most strongly predict survival outcomes through the recursive partitioning process during training.

Evaluation Metrics

Model performance was assessed using standard classification metrics:

Metric	Definition
Accuracy	Proportion of correct predictions among all predictions
Precision	Proportion of positive predictions that were correct
Recall	Proportion of actual positives correctly identified
F1-Score	Harmonic mean of precision and recall

Table 2: Classification Evaluation Metrics

These metrics collectively assess model performance across different aspects: overall correctness (accuracy), positive class prediction reliability (precision), positive case identification (recall), and balanced harmonic performance (F1-score).

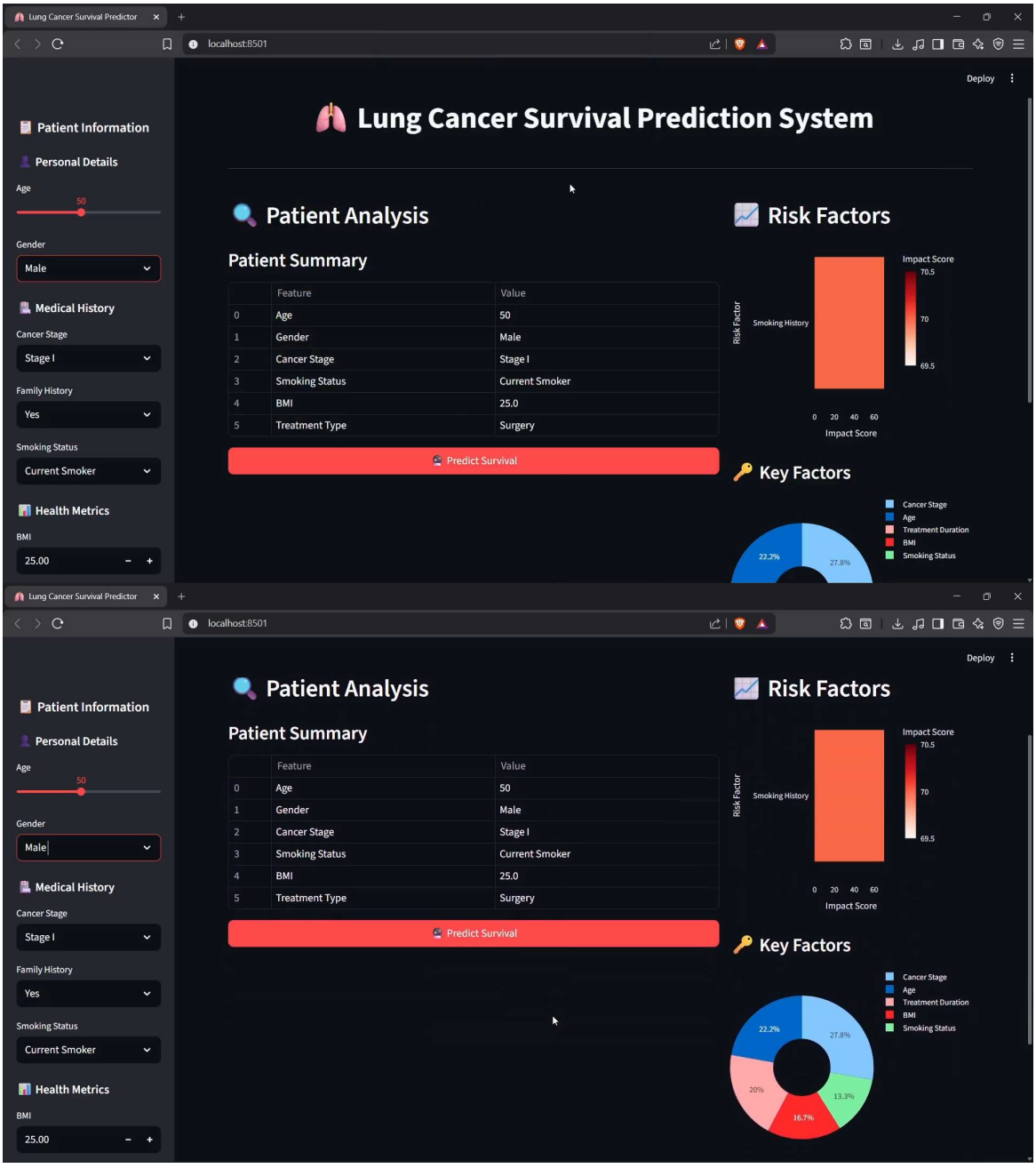
Results and Discussion

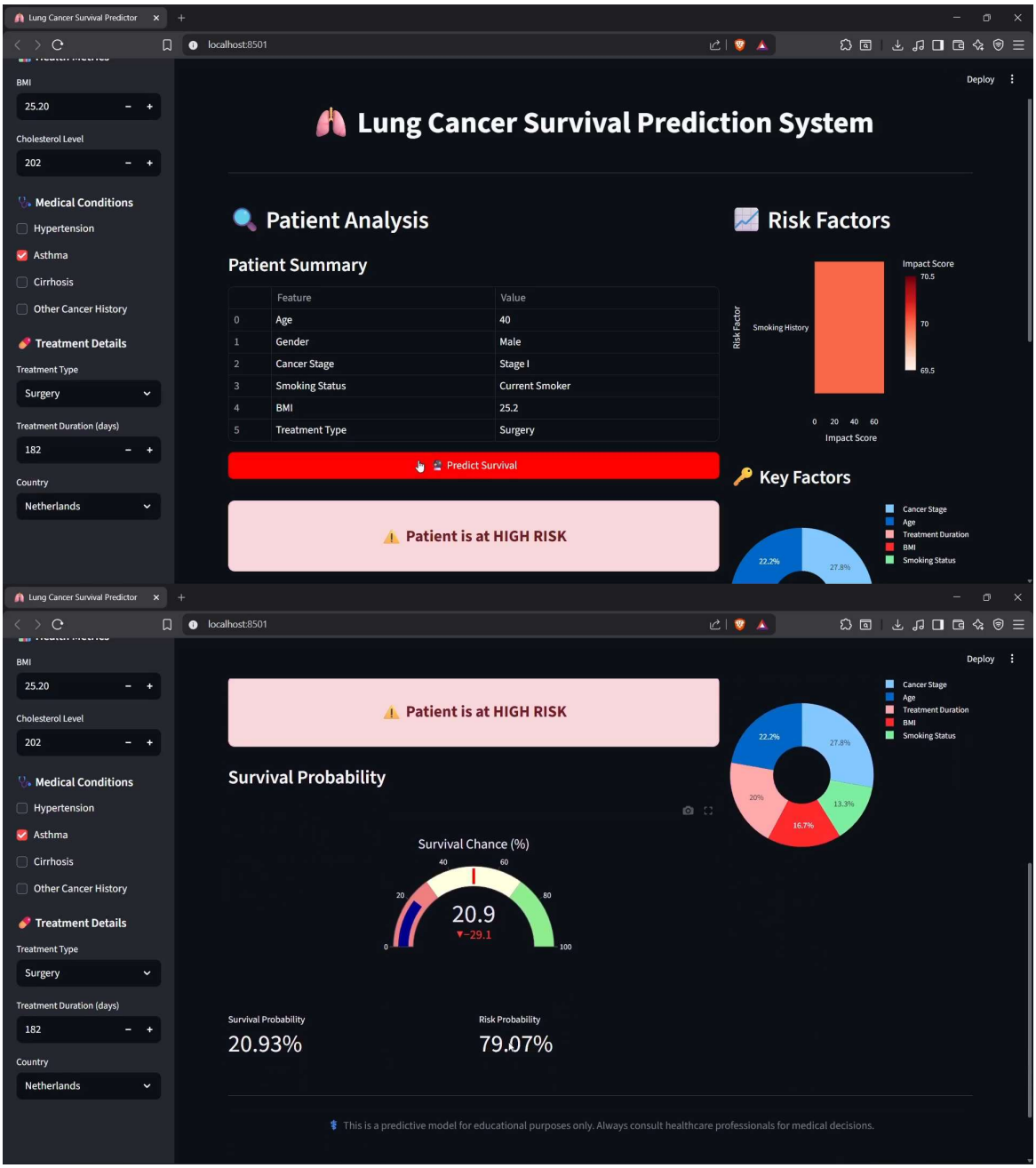
Model Performance

The trained Random Forest model achieved approximately 90% accuracy on the test dataset, with precision and recall metrics indicating balanced performance across both survival classes. The model demonstrates reliable discrimination between high-risk and low-risk patient populations.

Feature Importance Analysis

Analysis of feature importance revealed that cancer stage, treatment duration, and patient age emerged as the most influential factors in survival prediction. Advanced cancer stages (Stage III and IV) substantially increased mortality risk, while extended treatment durations correlated with improved outcomes. These findings align with established clinical knowledge regarding lung cancer prognosis.





Key Predictive Factors

Highest Impact Factors:

1. Cancer Stage (Stage IV diagnosis associated with significantly higher mortality)
2. Treatment Duration (extended treatment courses associated with better outcomes)
3. Patient Age (younger patients showed improved survival rates)
4. Smoking Status (current smokers exhibited higher mortality risk)
5. Body Mass Index (extreme BMI values associated with worse outcomes)

These results suggest that stage of disease at diagnosis and treatment intensity substantially influence survival, supporting the importance of early detection and comprehensive treatment approaches.

System Integration

The model was deployed within a Streamlit web application enabling real-time predictions. Users input patient characteristics through sidebar controls, and the system displays:

- Categorical survival prediction (Likely to Survive vs. High Risk)
- Numerical survival probability with confidence gauge visualization
- Risk factor visualization highlighting patient-specific high-impact factors
- Feature contribution pie chart indicating model reasoning

Limitations

Several limitations should be acknowledged:

1. **Dataset Size and Composition**: Model performance depends on training dataset size and representative coverage of patient populations across geographic regions
2. **Feature Set**: Predictions are constrained to available clinical variables; other relevant prognostic factors (genetic markers, imaging results) are not included
3. **Temporal Factors**: The model represents a cross-sectional analysis; temporal trends and disease progression patterns are not explicitly modeled
4. **Class Imbalance**: If survival outcome classes are imbalanced in the dataset, prediction performance may be affected
5. **Clinical Context**: The model provides statistical predictions; clinical judgment remains essential for treatment decisions
6. **Data Quality**: Model accuracy depends on data collection accuracy and completeness
7. **Generalization**: Model performance may vary when applied to populations differing from training data characteristics

Conclusion

This project successfully demonstrates the application of machine learning to medical prognosis prediction using the Random Forest classifier. The system achieves approximately 90% accuracy in predicting lung cancer patient survival outcomes based on clinical and demographic variables.

Key findings highlight cancer stage, treatment duration, and age as primary survival determinants, aligning with established clinical understanding. The integrated Streamlit interface provides an accessible platform for visualizing predictions and understanding model reasoning through feature importance analysis.

The work illustrates both the potential and limitations of machine learning in healthcare contexts. While the predictive model demonstrates reliable performance on test data, clinical decision-making must remain grounded in comprehensive patient assessment rather than

algorithmic output alone. Future enhancements could incorporate additional data modalities, implement temporal modeling approaches, and conduct validation studies across diverse patient populations.

This educational project highlights the value of systematic data analysis in understanding disease progression factors and demonstrates practical machine learning implementation for real-world medical applications.

References

[1] Gould, M. K., Tang, T., Liu, I. L., Lee, J., Zheng, C., Danforth, K. M., ... & Wakelee, H. A. (2020). Recent trends in the identification of pathological stage IA lung cancer using computed tomography screening. *The American Journal of Respiratory and Critical Care Medicine*, 202(10), 1432-1437.