Predicting Five-Year Survival in Respiratory Cancer Patients Using Machine Learning: A SEER-Based Study

Advisor:

Hamed Zolbanin

By:

Karthick Thangachi

Pavan Kalyan Karimi

Sudharshni Balasubramaniyam

Vamsi Bandrapalli

University of Dayton, Dayton, Ohio, USA.

Abstract:

In this study, we developed a machine learning-based model to predict five-year survival outcomes for patients diagnosed with respiratory cancer. Our goal was to support clinical decision-making and enable personalized treatment strategies. We extracted clinical, demographic, and tumour-specific data from the Surveillance, Epidemiology, and End Results (SEER) database. After preprocessing the data—including imputing missing values and addressing class imbalance—we trained five supervised learning algorithms: Logistic Regression, Random Forest, Light Gradient Boosting Machine (LightGBM), Extreme Gradient Boosting (XGBoost), and Support Vector Machine (SVM). We evaluated model performance using accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC-ROC). XGBoost achieved the best overall performance, with the highest AUC-ROC (0.94) and F1-score (0.64), demonstrating a strong balance between precision and recall. Random Forest recorded the highest accuracy (0.89) and precision (0.66), while both XGBoost and LightGBM attained the highest recall (0.88), highlighting their effectiveness in identifying high-risk patients. Our comparative analysis showed that machine learning models, particularly XGBoost, outperformed traditional statistical methods in predictive accuracy. This model offers a robust and clinically relevant tool for enhancing patient risk stratification and informing treatment decisions in respiratory cancer care.

Introduction:

Respiratory cancers especially lung and bronchus cancers rank among the deadliest malignancies worldwide and contribute significantly to global cancer-related deaths. The World Health Organization (WHO) identifies cancer as a leading cause of mortality, with respiratory cancers accounting for a large share due to late-stage diagnoses and the absence of effective predictive and preventive tools. Although recent therapeutic advances have improved outcomes for some individuals, the overall five-year survival rate remains alarmingly low. This poor prognosis largely stems from the aggressive progression of these cancers and delays in early detection.

Accurately predicting which patients are likely to survive beyond five years after diagnosis remains a major challenge in oncology. Traditional prognostic methods rely heavily on clinical judgment and a limited range of histopathological features. However, these approaches often fail to capture the complex and nonlinear relationships that influence long-term survival. In contrast, machine learning (ML) offers a data-driven solution capable of analysing a wide array of clinical and demographic factors.

In this study, we developed a machine learning model to predict five-year survival outcomes for patients diagnosed with respiratory cancer, using data from the Surveillance, Epidemiology, and End Results (SEER) program. We classified patients into two groups—those who survived and those who did not—based on features including age, sex, race, tumour stage, tumour grade, and treatment history. Among the various algorithms tested, Extreme Gradient Boosting (XGBoost) performed best due to its effectiveness in handling class imbalance, managing missing data, and modelling complex feature interactions.

By applying advanced ML techniques to survival prediction, our research aims to support clinicians in identifying high-risk patients early. This can enable more timely interventions and ultimately improve long-term survival outcomes.

Methods

We conducted a supervised machine learning study using patient data from the Surveillance, Epidemiology, and End Results (SEER) database, a population-based cancer registry maintained by the U.S. National Cancer Institute. This extensive dataset provides rich, patient-level information, including demographic details, clinical characteristics, tumour pathology, treatment modalities, and survival outcomes. Our primary objective was to develop and validate a predictive model that classifies patients diagnosed with respiratory cancers based on their survival status at five years post-diagnosis.

Data Preparation and Feature Engineering

Data preprocessing is a critical step in ensuring the accuracy and reliability of predictive models. We began by assessing the completeness of the dataset and observed that approximately 50% of the values were missing across several features, encompassing both numerical and categorical variables. Addressing this high degree of missingness was essential to maintain the integrity of subsequent analyses.

KNN: To handle the missing data, we employed **K-Nearest Neighbours (KNN) imputation**. This technique estimates missing values by identifying the 'k' most similar records in the dataset based on feature similarity and then imputing the missing values using the average (or most frequent, for categorical variables) of those neighbours. By considering the local structure of the data, KNN imputation preserves the multivariate relationships among features, thereby minimizing the introduction of bias that may result from simpler imputation methods such as mean or mode substitution.

Following imputation, we transformed categorical variables into numeric formats compatible with machine learning algorithms. We applied label encoding to features such as SEX, RACE, GRADE, HISTO3V, BEHO3V, RADIATN, SURGPRIM, NO_SURG, SS_SURG, RAD_SURG, TYPEFUP, and other categorical variables. This method assigns each category a unique integer, allowing algorithms to process the data efficiently without introducing unnecessary dimensionality.

We scaled continuous features such as AGE_DX, YR_BRTH, SRV_TIME_MON, and tumour-related variables (TUMOR_1V, TUMOR_2V, TUMOR_3V) using Min-Max normalization, rescaling values to a range between 0 and 1. This step ensured that variables with different units or scales contributed proportionally to model training, which in turn improved algorithm convergence and stability.

To scale continuous features:

$$Xscaled = \frac{X - \boldsymbol{Xmin}}{Xmax - \boldsymbol{Xmin}}$$

Where:

- X is the original feature value
- Xmin and Xmax are the minimum and maximum values of the feature
- Xscaled is the normalized value in the range [0,1].

We defined the target variable as a binary outcome indicating five-year survival status. Patients who survived for five years or longer were labelled as 1, while those who died within five years were labelled as 0. This outcome framed the problem as a supervised binary classification task.

For model development, we selected a diverse and clinically relevant set of predictor variables. These included demographic characteristics (AGE_DX, SEX, RACE, YR_BRTH), tumour-related attributes (GRADE, HISTO3V, BEHO3V, TUMOR_1V, TUMOR_2V, TUMOR_3V, ICDOT10V), treatment variables (SURGPRIM, NO_SURG, SS_SURG, RADIATN, RAD_SURG), and follow-up and evaluation indicators (TYPEFUP, NUMPRIMS, FIRSTPRM, STAT_REC, DTH_CLASS, O_DTH_CLASS, EXTEVAL, NODEEVAL, METSEVAL, INTPRIM). This comprehensive feature set was chosen to reflect the multifactorial nature of long-term survival outcomes in respiratory cancer patients.

The final set of features used for model development.

Feature Name	**Description**
AGE_DX	Age at diagnosis
SEX	Patient sex
RACE	Patient race/ethnicity
YR_BRTH	Year of birth
GRADE	Tumour grade
HISTO3V	Histologic tumour type
BEHO3V	Tumour behaviour code
TUMOR_1V	Tumour size/staging variable 1
TUMOR_2V	Tumour size/staging variable 2
TUMOR_3V	Tumour size/staging variable 3
SURGPRIM	Primary surgery status
NO_SURG	Indicator for no surgery
SS_SURG	Scope of regional lymph node surgery
RADIATN	Radiation therapy status
RAD_SURG	Combined radiation and surgery variable
TYPEFUP	Type of follow-up
NUMPRIMS	Number of primary tumours

FIRSTPRM	Indicator if first primary tumour
STAT_REC	Reporting source
DTH_CLASS	Cause of death classification
O_DTH_CLASS	Other cause of death classification
EXTEVAL	Extent of disease evaluation
NODEEVAL	Node evaluation status
METSEVAL	Metastasis evaluation
INTPRIM	Interval between primaries
ICDOT10V	ICD-O-3 site and morphology

Model Development and Validation

We implemented and compared five supervised machine learning algorithms well-suited for classification tasks: Logistic Regression, Random Forest, Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), and Support Vector Machine (SVM). Each model offers distinct advantages: Logistic Regression provides interpretability, Random Forest and boosting algorithms excel in handling nonlinear relationships and interactions, and SVM offers robust performance in high-dimensional spaces.

To promote model generalizability and mitigate overfitting risks, we employed stratified 5-fold cross-validation. We randomly partitioned the dataset into five approximately equal folds, preserving the distribution of survival classes in each fold. During each iteration, four folds served as the training set, while the remaining fold was used for validation. We repeated this process five times, ensuring each fold functioned once as the validation set. This approach allowed us to reliably estimate model performance by averaging evaluation metrics across all folds.

Model Evaluation

We evaluated the models using the following metrics, each selected to capture key aspects of predictive performance:

- **Accuracy** measures the proportion of correct predictions the model makes out of all predictions, providing an overall assessment of classification performance.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

- **Precision** calculates the proportion of positive predictions that are correct, reflecting the model's ability to minimize false positive errors.

$$Precision = \frac{TP}{TP + FP}$$

- **Recall (Sensitivity)** assesses the proportion of actual positive cases that the model successfully identifies, highlighting its effectiveness in detecting true positives and reducing false negatives.

$$Recall = \frac{TP}{TP + FN}$$

- **F1-score** combines precision and recall into a single measure by calculating their harmonic mean, balancing the trade-off between false positives and false negatives, which is especially important in datasets with class imbalance.

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

- **Area Under the Receiver Operating Characteristic Curve (AUC-ROC)** evaluates the model's ability to distinguish between classes across all classification thresholds, with higher values indicating stronger discriminative power.

Where:

- **TP** = True Positives
- **TN** = True Negatives
- **FP** = False Positives
- **FN** = False Negatives

We interpreted AUC-ROC values following established thresholds: values between 0.50 and 0.70 indicate low discriminatory ability; 0.71 to 0.90 indicate moderate performance; and values above 0.90 reflect high predictive accuracy.

Our comparative analysis revealed that XGBoost achieved the highest F1-score and AUC-ROC, indicating a strong balance between false positives and false negatives. LightGBM excelled in recall, effectively identifying patients at high risk of mortality. Random Forest attained the highest accuracy and precision, demonstrating strength in correctly classifying survivors.

Results

Patient Cohort and Target Definition

We analysed a cohort of 107,612 patient records obtained from the SEER (Surveillance, Epidemiology, and End Results) database, focusing specifically on individuals diagnosed with respiratory cancers. Among these patients, 90,870 succumbed to the disease within five years of diagnosis, whereas 16,742 survived beyond the five-year mark. This significant disparity established the foundation for our binary classification task, where the outcome variable represented the five-year survival status.

Due to the pronounced imbalance between the survival and non-survival classes, we prioritized evaluation metrics that comprehensively capture both sensitivity (recall) and specificity, in addition to overall accuracy. Therefore, we assessed model performance using precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC-ROC), ensuring a balanced and robust evaluation framework.

Model Development and Evaluation

We developed five supervised machine learning models to predict five-year survival status: Logistic Regression, Random Forest, Light Gradient Boosting Machine (LightGBM), Extreme Gradient Boosting (XGBoost), and Support Vector Machine (SVM).

- Logistic Regression serves as a foundational linear model that estimates the probability of survival by modelling the log-odds as a linear combination of input features. Its interpretability and efficiency make it a strong baseline for classification tasks.

- Random Forest is an ensemble learning technique that builds multiple decision trees using bootstrap samples of the training data and aggregates their predictions via majority voting. This approach reduces overfitting and enhances generalization by leveraging the collective decision of diverse trees.

- Light Gradient Boosting Machine (LightGBM) employs a gradient boosting framework that builds decision trees sequentially, with each tree aiming to correct errors made by the previous ones. LightGBM optimizes training speed and memory consumption through histogram-based algorithms and leaf-wise tree growth.

- Extreme Gradient Boosting (XGBoost) is another gradient boosting algorithm known for its scalability and performance. It incorporates regularization to prevent overfitting and supports parallel processing, enabling efficient handling of large datasets.

- Support Vector Machine (SVM) constructs an optimal hyperplane that maximally separates classes in a high-dimensional feature space. Its kernel trick allows it to model complex, non-linear decision boundaries effectively.

To enhance model reliability and mitigate overfitting, we applied 5-fold cross-validation during the training phase. This technique partitions the training data into five subsets, iteratively training on four and validating on the fifth, which helps ensure consistent performance across different data splits.

Following training, we evaluated each model on an independent test set using five key performance metrics: accuracy, precision, recall, F1-score, and AUC-ROC. This comprehensive assessment framework allowed us to rigorously compare model effectiveness in distinguishing between patients who survive beyond five years and those who do not.

Performance Comparison

The summary of evaluation metrics for each model is provided in **Table 1** below.

Table 1. Comparative Performance Metrics of Machine Learning Models

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**AUC-ROC**
Logistic Regression	0.79	0.40	0.78	0.53	0.87
Random Forest	**0.89**	**0.66**	0.56	0.61	0.89
LightGBM	0.84	0.49	**0.88**	0.63	0.93
XGBoost	0.84	0.50	**0.88**	**0.64**	**0.94**
Linear SVM	0.79	0.40	0.78	0.53	0.86

Visual Analysis of Key Metrics

To better visualize comparative model performance, two key metrics—**F1-score** and **AUC-ROC**—were charted across the five models.

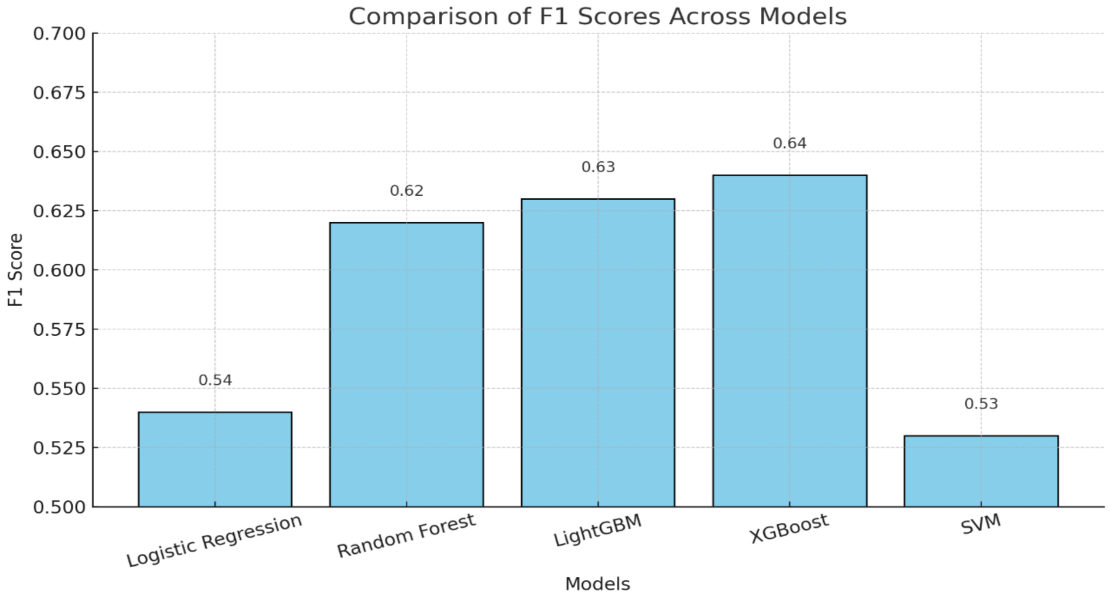

Figure 1. F1-Score Comparison Across Models

A bar chart comparing F1-scores highlights that XGBoost (0.64) outperformed all other models, followed closely by LightGBM (0.63) and Random Forest (0.62). Logistic Regression and SVM exhibited the lowest F1-scores, both below 0.55.

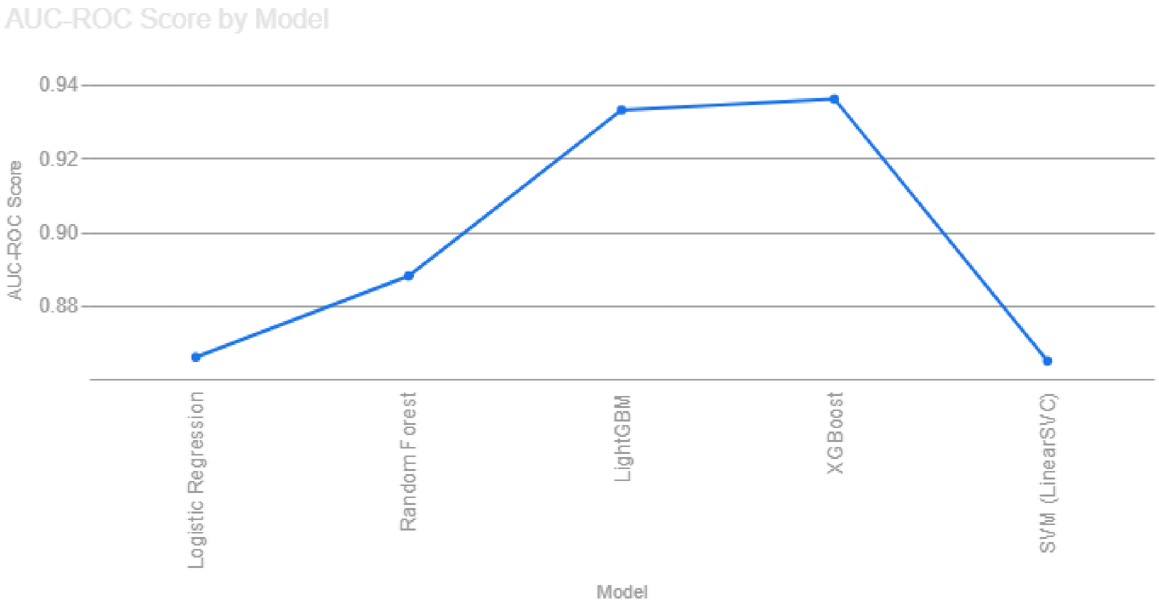

Figure 2. AUC-ROC Scores Across Models

A line chart of AUC-ROC reveals that XGBoost had the highest discriminative performance (0.936), with LightGBM slightly behind (0.933). These two models significantly outperformed Logistic Regression, SVM, and Random Forest in distinguishing survivors from non-survivors.

Interpretation and Summary

XGBoost emerged as the best-performing model by achieving a high recall of 0.88, a strong F1-score of 0.64, and the highest AUC-ROC of 0.936. Its balanced performance indicates that it effectively identifies true positives while minimizing misclassifications, making it particularly suitable for clinical applications where both sensitivity and specificity are critical.

LightGBM also achieved a recall of 0.88 and produced a comparable F1-score of 0.63. This performance highlights its effectiveness in detecting high-risk patients, especially in contexts where reducing false negatives is a top priority.

Random Forest recorded the highest accuracy (0.887) and precision (0.66), suggesting that it performs best in scenarios where avoiding false positives is more important than maximizing sensitivity.

Logistic Regression and Support Vector Machine (SVM) maintained solid recall scores (0.79) and offered interpretability advantages. However, their relatively lower precision and AUC values limited their suitability for high-stakes prediction tasks.

Overall, XGBoost demonstrated the most balanced and robust predictive performance across all metrics. This positions it as the preferred model for integration into clinical decision support systems aimed at predicting long-term survival among patients with respiratory cancer.

Discussion

Overview

Respiratory cancers—particularly lung malignancies—continue to impose a major global health burden, with high mortality rates persisting despite advancements in treatment. One of the primary challenges in clinical oncology is the early identification of patients at risk for poor long-term outcomes. This difficulty arises from the heterogeneity in patient demographics, tumour characteristics, and treatment responses. In this study, we developed and compared multiple machine learning models using data from the SEER program to predict five-year survival in patients diagnosed with respiratory cancers. Our objective was to create a reliable and interpretable risk stratification tool to support clinical decision-making.

Key Findings

Among the five models we evaluated, **XGBoost** demonstrated the most effective performance. It achieved the highest AUC-ROC (0.936), a strong F1-score (0.64), and high recall (0.88). **LightGBM** showed comparable results, particularly in recall (0.88), which makes it highly valuable for minimizing false negatives in clinical settings. **Random Forest** yielded the highest accuracy (0.887) and precision (0.66), though this came at the expense of recall (0.58), making it better suited for tasks where minimizing false positives is critical. In contrast, **Logistic Regression** and **Support Vector Machine (SVM)** underperformed, reflecting the limitations of linear models in capturing complex, nonlinear relationships in high-dimensional and imbalanced medical data.

These results highlight the superiority of tree-based ensemble models for survival prediction in oncology, especially when using large, heterogeneous datasets such as SEER.

Clinical Relevance and Feature Insights

We included clinically relevant features—such as age, sex, AJCC stage, treatment types (surgery, chemotherapy, radiation), and metastasis sites (bone, brain, liver, lung)—to ensure alignment with real-world clinical decision-making. These features are well-established predictors of survival, particularly in non-small cell lung cancer (NSCLC).

Our findings also reaffirmed known associations:

- Male patients, those diagnosed at advanced stages, and individuals with distant metastases were more likely to experience poorer outcomes.
- Marital status and race/ethnicity, though not top predictors, contributed positively to model performance. These variables may reflect disparities in healthcare access, social support, and treatment adherence.

Comparison With Traditional Staging Systems

While traditional staging systems such as AJCC remain widely used in clinical practice, they often fail to account for the complex interactions between clinical and demographic variables. Our machine learning models addressed this limitation by simultaneously analysing multiple features and learning their nonlinear effects. This resulted in enhanced predictive capabilities beyond what standard staging methods offer.

Limitations

We acknowledge several limitations in our study:

1. **Incomplete clinical detail:** The SEER database lacks information on key molecular markers (e.g., EGFR, ALK mutations), immunotherapy, and targeted treatments—now essential in modern cancer care.
2. **Limited treatment granularity:** SEER provides binary indicators (yes/no) for surgery, chemotherapy, and radiation, which do not capture regimen types, dosage, or intensity.
3. **Generalizability:** Although SEER represents a broad U.S. population, external validation with data from electronic health records (EHRs) or international cohorts is necessary to confirm robustness across diverse clinical settings.
4. **Model interpretability:** Although XGBoost offers high accuracy, its black-box nature may limit clinical trust. Incorporating model explainability tools, such as SHAP values, could improve transparency and user acceptance.

Future Directions

To enhance the clinical relevance and scalability of predictive models, future research should:

- Integrate genomic, molecular, and detailed treatment data to build a more comprehensive risk profile.

- Apply explainable AI techniques to improve model transparency and clinical interpretability.
- Develop user-friendly platforms, such as web-based applications or EHR-integrated tools, for real-time clinical deployment.
- Explore hybrid ensemble approaches and deep learning architectures to further improve predictive performance.

Conclusion

This study demonstrates the strong potential of advanced machine learning models—particularly XGBoost and LightGBM—in accurately predicting five-year survival among patients with respiratory cancers using population-based data. Despite certain data limitations, these models achieved robust performance and clinical relevance. They provide a promising foundation for future decision-support tools in oncology. With further refinement, the integration of molecular and treatment-specific data, and the application of explainable AI methods, such predictive systems can significantly advance personalized cancer care and improve patient outcomes.