Prognostic Factors Affecting Survival Outcomes in Lung Adenocarcinoma

Abhay Rastogi, Chetan Elenki

Abstract

Lung adenocarcinoma (LUAD) exhibits significant heterogeneity in survival outcomes due to clinical, demographic, and molecular factors. This study analyzed an integrated dataset of 989 LUAD patients from TCGA-LUAD, OncoSG, and SHERLOCK lung cancer studies. Supervised models achieved moderate performance (test accuracy 70.5%, ROC-AUC 0.69 for Naïve Bayes; C-index 0.734 for CoxNet), identifying age, tumor stage, and TP53 mutations as the most significant predictors of survival. Unsupervised clustering revealed six distinct subtypes including the Asian non-smoker *EGFR*-positive and Caucasian smoker *KRAS*-positive phenotypes. Age emerged as a dominant prognostic factor, with younger patients showing better survival despite adverse molecular profiles, challenging conventional staging systems. Kaplan-Meier analysis confirmed significant survival differences based on smoking status, tumor stage, and TP53 mutations. In contrast, SHAP analysis highlighted the critical impact of age, tumor stage, and ethnicity-linked mutation patterns. Gaussian Mixture Model clustering further defined three prognostically distinct groups, with survival ranging from 53.7 months to over 167 months. These findings underscore the complexity of LUAD heterogeneity and highlight the need for personalized risk stratification.

Introduction

Lung adenocarcinoma represents the most common histological subtype of lung cancer, accounting for over 40% of all cases. Despite advancements in targeted therapies and early detection strategies, five-year survival rates remain poor, ranging from 20-30% in early-stage disease to less than 10% in advanced stages. Clinical prognostication of LUAD is based on the Tumor Node Metastasis (TNM) system, which examines the anatomical extent to which the disease has spread. However, sociodemographic disparities in lung cancer survival are well-documented **(1)**. Factors such as race, ethnicity, and underlying genomic differences influencing outcomes have significantly impacted survival outcomes. Hence, these factors must be considered in addition to anatomical staging to provide better survival outcome estimates for individual patients.

A systematic analysis of 32 published lung cancer prognostic models **(2)** showed that motivations behind tool development were identifying circumstances relevant to specific patient populations or prioritizing the inclusion of emerging factors contributing to oncogenesis, such as new environmental exposures. Similarly, this study was motivated by the need to improve prognostic accuracy by including factors such as age, ethnicity, and the presence of key driver mutations in addition to tumor stage and gain insights about which factors should be prioritized in patient subgroups. This comprehensive risk-assessment approach could yield better stratification of patients for cancer screening, treatment prioritization, and tailored disease management.

Background

Lung adenocarcinoma (LUAD) shows striking survival differences based on demographic and epidemiological patterns. While historically linked to smoking, East Asian LUAD patients tend to be younger, non-smoking females. Extensive genomic clinical studies such as The Cancer Genome Atlas (TCGA) demonstrated that patients with targetable mutations (*EGFR, KRAS, ALK-ROS*) have better prognoses than TP53-mutated LUAD **(3)**. The SHERLOCK study **(4)** of 232 never-smoker LUAD patients revealed three distinct subtypes, with the *EGFR*-mutated subtype characterized by longer telomeres and lower genomic instability showing better survival outcomes; the Singaporean OncoSG study **(5)** of 305 Chinese-ancestry LUAD patients showed East-Asians have more stable genomes with fewer mutations compared to Caucasians, particularly among smokers. The Taiwanese TALENT trial **(6)** further showed over 50% *EGFR* positivity in East Asian patients (as compared to 10-15% in Caucasians) and better overall survival, even after adjusting for known prognostic factors.

Previous machine learning research on cancer survival predictions demonstrated promising results across multiple types. Using gene expression data for lung adenocarcinoma, Chen *et al.* **(7)** developed artificial neural networks with 83% accuracy and identified key prognostic genes (*LCK* and *ERBB2*). Using cancer registry data for breast cancer, Gupta *et al.* **(8)** achieved AUCs of 0.87 at 6 months and 0.76 at 24 months. Support Vector Machines (SVMs) performed exceptionally well in their study due to their robustness in handling high-dimensional data. In their review, Kourou *et al*. **(9)** showed that integrating clinical and molecular data improved prediction accuracy across different methods, particularly for lung, breast, and oral cancers, with AUCs ranging from 0.71 to 0.97. Their analysis revealed that ensemble and deep learning approaches outperformed traditional statistical methods for LUAD survival prediction. However, these studies also highlighted essential challenges, including issues with data quality, the need for external validation, the importance of feature selection, and the difficulty of model generalizability across different patient populations.

Methodology and Results

Datasets and exploratory analysis

Clinical and genomic data from three lung adenocarcinoma clinical studies, namely TCGA-LUAD (n = 494), OncoSG (n = 305), and SHERLOCK (n = 232), were obtained using the cBioPortal web API. Features included sex, race, tumor stage, age, three driver mutations (*TP53, EGFR, KRAS*), and survival status and time as targets (Appendix 1). The exploratory analysis of the combined LUAD dataset (n = 989) identified a class imbalance, with the "deceased" group being the minority class (33%). The mean age at diagnosis was 64.7 years, with a mean overall survival of 38.7 months. Cramer's V analysis showed weak correlations between features. Kaplan-Meier (KM) plots indicate significant survival differences based on smoking status, tumor stage, and TP53 mutations, with advanced stages and *TP53* mutations associated with worse prognosis.

Preprocessing and analysis pipeline

Data preprocessing involved cleaning, encoding, and handling class imbalance. Excluding instances with missing values (<7%) resulted in better model performance than data imputation. Categorical and numerical features were one-hot-encoded and scaled using sckit-learn OneHotEncoder and StandardScaler modules respectively. To address class imbalance in the binary classification task, SMOTE (Synthetic Minority Oversampling Technique) from the imlearn library was applied. Seven supervised classification models and two regression models were trained (Table 1). Stratified 10-fold cross-validation was implemented with scikit-learn StratifiedKFold to balance evaluation across folds. Bayesian optimization-based hyperparameter tuning was performed using the Optuna library.

Supervised Classification and Regression models

Among supervised classification models, naive bayes achieved the highest test accuracy of 70.5% and a cross-validation accuracy of 70.0%. Bagging classifier and SVM also performed well, with test accuracies of 69.9% and 71.6%, respectively. The random forest and logistic regression models had balanced precision and recall, making them suitable for imbalanced datasets. XGBoost and CatBoost achieved moderate performance with test accuracies of 64.8% and 63.1%, respectively, but struggled with minority class precision. ROC-AUC scores ranged from 0.62 to 0.69, with naive bayes and bagging classifier achieving the highest at 0.69 and 0.68, respectively.

CoxNet achieved the highest concordance index (C-index) for regression models of 0.734, followed by random survival forest at 0.714. These results indicate that CoxNet was more effective in predicting survival times, while random survival forest provided additional interpretability with survival curve predictions. Overall, age, *TP53* mutations, and tumor stage were consistently identified as the most influential predictors across all models, validated by SHAP analysis.

Table 1: Performance and best hyperparameters for classification and regression models

Model	**Best Hyperparameters**	**Test Accuracy / C-index**	**Cross-validation accuracy**	**Precision (0/1)**	**Recall (0/1)**	**F1-Score (0/1)**	**ROC-AUC**
Random Forest	n_estimators: 156, max_depth: 9	0.676	0.69 ± 0.03	0.72 / 0.50	0.84 / 0.33	0.78 / 0.40	0.67
XGBoost	n_estimators: 222, learning_rate: 0.29, max_depth: 16	0.648	0.67 ± 0.04	0.75 / 0.46	0.71 / 0.51	0.73 / 0.48	0.62
SVM	C: 0.11, kernel: poly	0.716	0.71 ± 0.02	0.74 / 0.60	0.88 / 0.37	0.81 / 0.46	0.65
CatBoost	n_estimators: 238, learning_rate: 0.25, depth: 9	0.631	0.65 ± 0.05	0.72 / 0.43	0.73 / 0.42	0.73 / 0.42	0.66
Logistic Regression	C: 4.75	0.648	0.66 ± 0.03	0.77 / 0.46	0.68 / 0.58	0.72 / 0.52	0.68
Bagging Classifier	n_estimators: 131, max_depth: 5	0.699	0.68 ± 0.04	0.75 / 0.55	0.84 / 0.40	0.79 / 0.46	0.68
Naive Bayes	alpha: 0.43	0.705	0.70 ± 0.02	0.75 / 0.56	0.85 / 0.40	0.80 / 0.47	0.69
Coxnet	l1_ratio: 0.9	0.734 (C-index)	N/A	N/A	N/A	N/A	N/A
Random Survival Forest	n_estimators: 266, min_samples_split: 18, min_samples_leaf: 7	0.714 (C-index)	N/A	N/A	N/A	N/A	N/A

*0: majority class (living); 1: minority class (deceased)

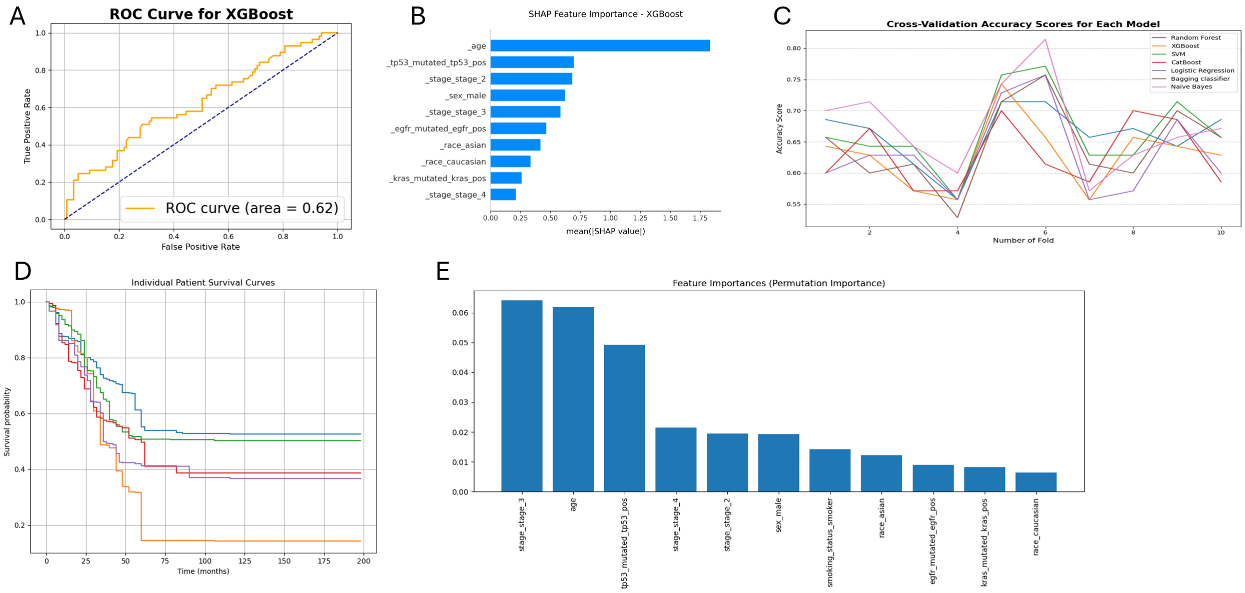

Fig 1. Supervised classification and regression models for combined LUAD dataset. (A) Receiver Operating Characteristic (ROC) curve illustrating the performance of the XGBoost classifier in predicting survival outcomes. (B) Summary plot displaying the SHAP values for features used in the XGBoost model. (C) Line plot showing cross-validation accuracy scores for different machine learning models across ten stratified validation folds. (D) Kaplan- Meier survival curves for five individual patients, derived from random survival forest supervised regression model. (E) Bar chart depicting the importance of permutation-based features across the random survival forest regression model.

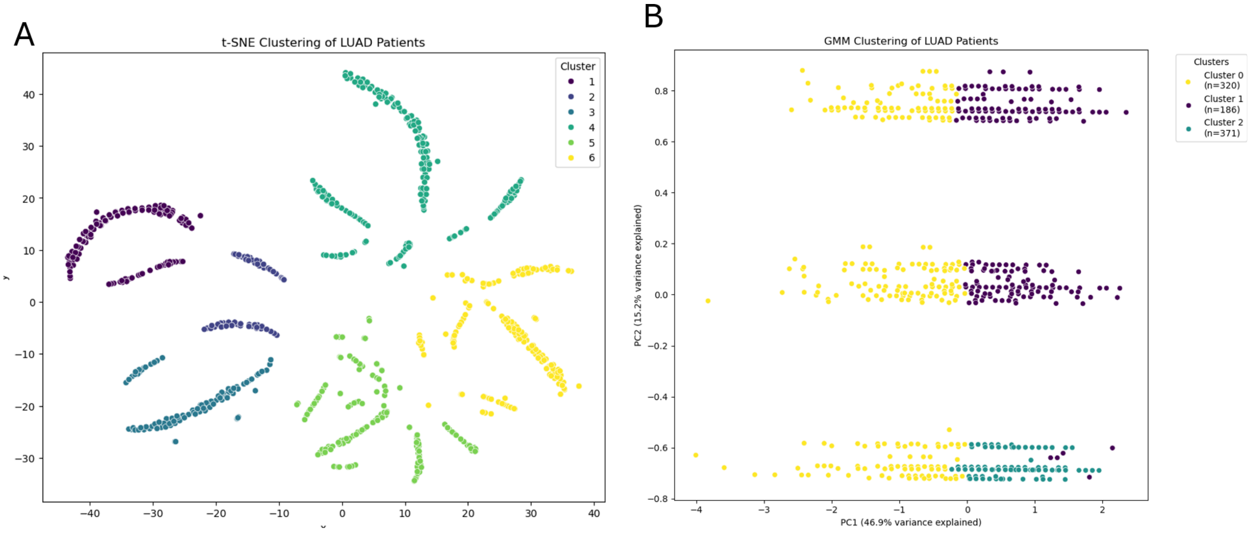

Fig 2. Unsupervised clustering for combined LUAD dataset. (A) t-distributed Stochastic Neighbor Embedding (t-SNE) visualization of LUAD patient clusters based on clinical and molecular features. (B) Gaussian Mixture Model (GMM) clustering of LUAD patients using the first two principal components (PC1 and PC2)

Unsupervised models

To identify natural subgroups within the LUAD patient population, t-SNE dimensionality reduction followed by hierarchical clustering was applied to the pre-processed dataset. Bayesian optimization with Optuna was used to tune the t-SNE parameters, yielding optimal values for perplexity (25.45) and learning rate (404.55), indicating moderate cluster separation. Six distinct LUAD subtypes were identified, with the two most significant clusters representing established molecular-clinical phenotypes: Cluster 4 (24.4%) captured the "Classic Caucasian Non-Smoker LUAD" subtype with a high EGFR mutation rate (36.9%) and predominantly early-stage disease. In contrast, Cluster 6 (20.5%) represented the "East Asian Never-Smoker LUAD" subtype with the highest *EGFR* mutation rate (59.4%) and lowest KRAS rate (3.3%). The remaining clusters demonstrated clear ethnic and smoking-based stratification: three clusters (1,2,3; totaling 35.1%) comprised Caucasian smokers with high *TP53* (>50%) and *KRAS* mutations, while Cluster 5 (20.0%) represented a unique mixed Asian/African American population with predominantly smoking history (88.6%) and balanced mutation profiles.

The Gaussian Mixture Model identified three clinically distinct clusters with significant prognostic implications ($p < 0.0001$). The largest cluster (Cluster 2, 42.3%) exhibited the best survival outcomes (infinite median) despite having the highest TP53 mutation rate (48.8%), characterized by younger age (median 57) and balanced ethnic/smoking distributions. In contrast, Cluster 0 (36.5%) showed the poorest survival (median 53.7 months), comprising predominantly older smokers (median age 72) with high TP53 mutations (43.1%). Cluster 1 (21.2%) represented a pure non-smoking population with the highest EGFR mutation rate (46.8%) and intermediate survival (167 months). Notably, age emerged as a more substantial prognostic factor than mutation status or disease stage, with younger patients showing better survival despite adverse molecular features.

DBSCAN clustering analysis was performed and has essentially failed to find meaningful clusters.

Implementation

The dataset and code used for analysis are accessible here: https://github.com/abhayr20/bmi6015_project

Discussion

Our supervised learning models achieved moderate predictive performance, with age, TP53 mutations, and tumor stage emerging as the most consistent predictors across all models. A major limiting factor was class imbalance – only a third of all instances belonged to the minority class (diseased patients). To address the class imbalance in the dataset, the Synthetic Minority Oversampling Technique (SMOTE) was employed. SMOTE generates synthetic examples for the minority class by interpolating existing examples. This helped improve the recall performance for the minority class. Additionally, stratified 10-fold cross-validation was used to ensure balanced evaluation across folds. Imputing missing data reduced the performance of supervised models. Removing examples with missing features resulted in the loss of about 7% of the dataset. Still, it improved overall model performance while maintaining the proportional representation of features in the original combined dataset.

The most striking findings came from our unsupervised analyses, which revealed distinct molecular subtypes and continuous spectra of disease characteristics. t-SNE analysis identified six distinct clusters that strongly aligned with known biological patterns, particularly the established dichotomy between Asian non-smoker *EGFR*-positive and Caucasian smoker *KRAS*-positive subtypes. GMM clustering revealed three broader groups with significant prognostic differences ($p < 0.0001$), highlighting age as a surprisingly strong predictor. Notably, younger patients showed better survival despite adverse molecular features, suggesting age-specific disease mechanisms that warrant further investigation.

Our findings suggest that age should be given more significant consideration in treatment decisions, as it appears to be a more substantial prognostic factor than previously recognized. The analysis also indicates that ethnic-specific molecular testing strategies may be beneficial, given the strong association between ethnicity and mutation patterns. Through this analysis, we highlight the need for more nuanced risk stratification approaches that consider the continuous nature of disease characteristics rather than rely solely on discrete classifications.

Study limitations include the retrospective nature of the data and potential selection bias in the study populations. The absence of critical clinical, socioeconomic, and treatment-related factors—such as comorbidities, healthcare access, a fraction of genome mutated, and treatment details (such as surgery, chemotherapy, or radiation therapy) likely restricted the scope of the analyses. Furthermore, the dataset lacked comprehensive genomic and epigenetic profiling, including methylation data, and instead focused only on three driver mutations. Future work should focus on deriving datasets with complete genomic and epigenetic profiles and key prognostic features to enhance the predictive power of the models.

Conclusion

Our analysis suggests that traditional prognostic frameworks might be oversimplifying the complex biological reality of LUAD. The discovery that age transcends established molecular risk factors in predicting survival and identifying distinct ethnic-molecular subtypes existing along a continuous spectrum rather than discrete entities suggests the need to reconsider how we fundamentally stratify and treat LUAD patients. These findings indicate a shift from the current categorical classification systems toward more nuanced, personalized approaches considering the interplay between demographic, clinical, and molecular characteristics in treatment decision-making.

REFERENCES

1. Brouwer, A. F., Engle, J. M., Jeon, J. & Meza, R. Sociodemographic Survival Disparities for Lung Cancer in the United States, 2000-2016. *JNCI: J. Natl. Cancer Inst.* **114**, 1492–1500 (2022).

2. Mahar, A. L. *et al.* Refining Prognosis in Lung Cancer A Report on the Quality and Relevance of Clinical Prognostic Tools. *J. Thorac. Oncol.* **10**, 1576–1589 (2015).

3. Collisson, E. A. *et al.* Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543–550 (2014).

4. Zhang, T. *et al.* Genomic and evolutionary classification of lung cancer in non-smokers. *Nat. Genet.* **53**, 1348–1359 (2021).

5. Chen, J. *et al.* Genomic landscape of lung adenocarcinoma in East Asians. *Nat. Genet.* **52**, 177–186 (2020).

6. Chang, G.-C. *et al.* Low-dose CT screening among never-smokers with or without a family history of lung cancer in Taiwan: a prospective cohort study. *Lancet Respir. Med.* **12**, 141–152 (2024).

7. Chen, Y.-C., Ke, W.-C. & Chiu, H.-W. Risk classification of cancer survival using ANN with gene expression data from multiple laboratories. *Comput. Biol. Med.* **48**, 1–7 (2014).

8. Gupta, S. *et al.* Machine-learning prediction of cancer survival: a retrospective study using electronic administrative records and a cancer registry. *BMJ Open* **4**, e004007 (2014).

9. Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V. & Fotiadis, D. I. Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J.* **13**, 8–17 (2015).

APPENDIX

Appendix 1: Table of Features in Combined LUAD dataset (n=989)

Feature	Type	Description	Categories/Range
Sex	Categorical	The biological sex of the patient.	Male, Female
Race	Categorical	Self-reported race of the patient.	Caucasian, Asian, African American, Other
Tumor Stage	Categorical	Stage of LUAD at diagnosis.	Stage 1, Stage 2, Stage 3, Stage 4
Age	Numerical	Age of the patient at the time of diagnosis.	30–90 years (mean: ~64 years)
TP53 Mutation	Binary	Presence or absence of TP53 mutation in the tumor.	TP53_Pos, TP53_Neg
EGFR Mutation	Binary	Presence or absence of EGFR mutation in the tumor.	EGFR_Pos, EGFR_Neg
KRAS Mutation	Binary	Presence or absence of KRAS mutation in the tumor.	KRAS_Pos, KRAS_Neg
Survival Status	Binary	Whether the patient was alive or deceased at the end of the study.	Living, Deceased
Survival Time	Numerical	Overall survival time is in months from diagnosis until death or the end of the study.	0–200+ months (mean: ~39 months)