Investigation of the Effectiveness of Machine Learning Models in Lung Cancer Survival Prediction Using Synthetic Patients Data

Richard Gamah Teye

ABSTRACT

Lung cancer remains a leading cause of cancer-related deaths globally. Accurate prediction of survival is crucial for effective patient management, treatment planning, and resource allocation. Traditional prognostic factors, such as stage and histology, offer valuable insights, but they often fall short in providing the personalized information needed for precision medicine. The rapidly evolving field of precision prevention and medicine emphasizes tailoring interventions based on individual risk profiles, which importantly include genetic predispositions. This research investigates the application of various machine learning models to predict lung cancer survival. A key focus is identifying the factors most influential in these predictions, as they can inform precision prevention strategies and potentially highlight indicators of underlying genetic vulnerabilities impacting patient outcomes. By analyzing a comprehensive dataset of clinical and lifestyle factors, we aim to develop predictive models capable of contributing to more targeted screening, earlier detection, and personalized treatment approaches. Ultimately, this work seeks to improve survival rates and advance the field of precision oncology.

INTRODUCTION

Lung cancer stands as a formidable global health challenge, consistently ranking among the most prevalent cancers and the primary cause of cancer-related mortality worldwide [1][2][3]. Despite significant advancements in diagnosis and treatment modalities, the prognosis for lung cancer patients remains highly variable [4][5]. This variability is influenced by a complex interplay of factors, including tumor characteristics, the patient's overall health status, environmental exposures, and their unique genetic makeup [6][7][8][9][10][11]. The inherent heterogeneity of the disease underscores the critical need for accurate and personalized survival prediction [12][13].

Precise estimation of a patient's survival probability is paramount in the clinical management of lung cancer [14]. It directly informs crucial decisions regarding treatment strategies, dictating the intensity of therapy and guiding the selection of specific modalities such as surgery, chemotherapy, radiation therapy, and targeted therapies [15][16]. Furthermore, accurate survival prediction facilitates realistic patient counseling, helps set appropriate expectations, and supports shared decision-making processes between clinicians and patients [17][18]. From a broader public health perspective, reliable survival prognostics are essential for effective healthcare resource allocation and for the design of more targeted screening and prevention programs [19].

The paradigm of precision prevention and medicine is fundamentally transforming healthcare by advocating for the tailoring of interventions to individuals based on their unique biological and environmental characteristics [20][21][22]. In the context of lung cancer, this involves moving beyond broad, generalized risk categories [22][23]. The goal is to identify individuals at genuinely higher risk based on a nuanced combination of environmental exposures, lifestyle factors, and, crucially, genetic vulnerabilities. Understanding precisely how specific genetic variations influence an individual's susceptibility to developing lung cancer, the speed and nature of its progression, and their

response to various treatments is fundamental to developing truly personalized prevention and treatment strategies [22][24].

The primary objective of this research work is to investigate the effectiveness of various machine learning models in predicting one-year survival in lung cancer patients. This utilizes a synthetic dataset encompassing a diverse range of demographic, lifestyle, clinical, and treatment information. By applying and rigorously evaluating models such as Logistic Regression, Decision Trees, Random Forests, Gradient Boosting, and Neural Networks, it aims not only to predict survival outcomes with improved accuracy but also to identify the most influential factors driving these predictions. These influential factors, particularly when considered in the context of known biological pathways and established gene-environment interactions, can provide valuable insights. They may serve as potential indicators of underlying genetic vulnerabilities and inform the development of more precise prevention, early detection, and personalized treatment approaches for lung cancer. This work contributes to the broader goal of leveraging data science and machine learning to advance the field of precision oncology and ultimately improve patient outcomes on a personalized level.

RELATED WORKS

Predicting survival in lung cancer has long been a significant area of research, fundamentally driven by the imperative for improved prognostication and the development of personalized treatment strategies [13][17][18]. Over the years, numerous studies have meticulously explored the utility of various clinical, demographic, and environmental factors in predicting patient outcomes [25].

Early research efforts often concentrated on the analysis of established prognostic factors. These included widely recognized indicators such as the clinical stage of the cancer, its specific histology type, and the patient's overall performance status [26]. Landmark studies, notably those conducted by the International Association for the Study of Lung Cancer (IASLC), have consistently underscored the paramount importance of accurate staging which specifically utilizing the TNM classification system, in determining prognosis across the spectrum of different histology types, including adenocarcinoma, squamous cell carcinoma, and small cell carcinoma [27]. Furthermore, tumor size has been universally acknowledged and validated as a key predictor of survival outcomes, with a clear consensus that larger tumors are generally associated with poorer prognoses [28].

Demographic factors, prominently including age and gender, have also been subject to extensive investigation [29]. While advanced age is an established risk factor for the development of lung cancer, its independent prognostic value specifically in the context of survival prediction has yielded somewhat mixed and occasionally contradictory results across different studies [30]. Its impact often appears to vary significantly depending on the specific characteristics of the patient cohort under study and the nature of the treatment received [30]. Observed differences in survival rates between genders in certain populations have also been documented [31].

Lifestyle choices and environmental exposures, most notably smoking status, have been studied for their impact on both the risk of developing lung cancer and subsequent survival outcomes [32]. Smoking is unequivocally the single leading cause of lung cancer, and robust evidence demonstrates that continued smoking after a diagnosis is associated with a significantly poorer prognosis [33]. Environmental factors, such as

chronic exposure to air pollution, including fine particulate matter and various industrial pollutants, have also been firmly linked to an increased incidence of lung cancer [34][32]. While their direct and independent influence on survival prediction in routine clinical settings is an area where findings are less consistently established, it remains a crucial consideration [34]. Biomass fuel use, a significant and pervasive source of indoor air pollution in many regions globally, has been clearly identified as a substantial risk factor for lung cancer, and its specific influence on long-term survival is currently an important area of ongoing investigation [35]. A documented family history of lung cancer represents another critical factor, strongly suggesting a potential underlying genetic predisposition that could impact both an individual's susceptibility to the disease and their subsequent prognosis [36]. Dietary habits have similarly been explored in research, with some studies suggesting that specific dietary patterns may be associated with altered lung cancer risk and potentially influence survival outcomes [37]. This influence is likely mediated through complex interactions involving both genetic factors and environmental exposures [36][37].

Furthermore, the specific type of treatment received which includes encompassing modalities such as surgery, chemotherapy, radiation therapy, targeted therapy, and palliative care, is an important determinant of survival [38][39]. Treatment selection is typically based on a comprehensive assessment combining cancer stage, histology with or without the patient's overall health status [38]. The characteristics of the healthcare facility where treatment is administered, such as whether it is a private hospital, government institution, or medical college, can also play a notable role [40]. This is potentially reflective of differences in available resources, specialized expertise, access to advanced technologies, and adherence to established best practice treatment protocols, all of which could indirectly impact patient outcomes [41]. The specific symptoms presenting at the time of diagnosis can also furnish valuable prognostic information, as certain symptom profiles may indicate more advanced or aggressive disease progression [41][40].

In more recent years, machine learning techniques have been increasingly and successfully applied to the domain of lung cancer survival prediction [12][13][42]. These methods offer the capability to integrate diverse and complex factors and to identify intricate patterns and non-linear relationships that may not be readily apparent or detectable through traditional statistical methodologies [43]. Studies have effectively utilized a variety of models, including logistic regression, support vector machines, random forests, and artificial neural networks. This research has consistently demonstrated promising results in improving predictive accuracy when compared to conventional prognostic approaches [43].

Despite these significant and valuable advances, a notable gap persists in the field: the comprehensive integration of the understanding of genetic vulnerabilities into predictive models for lung cancer survival, particularly when utilizing readily available clinical and demographic data. While direct genetic sequencing provides information on specific mutations and genetic alterations, it is not universally available or feasible in all clinical settings. This research endeavors to explore whether standard clinical, demographic, and environmental factors, when analyzed through advanced machine learning techniques, can effectively serve as valuable indicators or correlates of underlying genetic influences on survival outcomes. By identifying which of these accessible factors are most important and influential in the predictive models, we can gain crucial insights into underlying biological pathways or patient characteristics that may be linked to genetic predispositions. This understanding can, in

turn, inform the development of more precise and targeted prevention strategies and help identify individuals who might benefit most significantly from targeted genetic screening or personalized interventions. This work seeks to bridge the existing gap between readily available clinical data and the immense promise of precision oncology by leveraging the power of machine learning to uncover subtle, yet significant, patterns relevant to genetic vulnerability and ultimately enhance personalized survival prediction.

METHODS

Dataset

The dataset utilized in this study is a synthetic lung cancer dataset openly available from Kaggle, specifically titled "Large Synthetic Lung Cancer Dataset - Bangladesh Perspective." This dataset was meticulously synthesized to realistically model a diverse range of clinical and lifestyle factors pertinent to lung cancer development and survival outcomes. Its primary purpose is to support research endeavors in the vital area of survival prediction. The dataset comprises information on a total of 5000 distinct patients. Each patient is characterized by 17 features, providing a comprehensive profile. These features include essential patient demographics (Age, Gender, Residence), various identified risk factors and lifestyle indicators (Smoking Status, Air Pollution Exposure, Biomass Fuel Use, Factory Exposure, Family History, Diet Habit), key primary symptoms commonly associated with lung cancer (Hemoptysis, Chest Pain, Fatigue & Weakness, Chronic Cough, Unexplained Weight Loss), crucial tumor characteristics and clinical features (Tumor Size in millimeters, Histology Type, Cancer Stage) and relevant information concerning treatment received and the healthcare facility where treatment was provided (Treatment Received, Hospital Type). The dataset includes a binary target variable, designated "Survival (Binary)", which clearly indicates whether a patient successfully survived for at least one year following their initial diagnosis.

Data Loading and Preprocessing

The dataset was loaded into a pandas DataFrame directly from its designated KaggleHub path. The initial structure of the DataFrame contained a mix of both numerical and categorical features. For the subsequent application of machine learning models, the *'Patient_ID'* column was excluded from the feature set, as it serves solely as a unique identifier and holds no predictive value. The remaining features were designated as the input features (represented by X), and the *'Survival_1_Year*' column was set as the target variable (represented by y).

Comprehensive data preprocessing was essential and involved the systematic encoding of both the categorical features and the target variable into a numerical format suitable for various machine learning algorithms. One-hot encoding was applied to all nominal categorical features present in the dataset. This included *'Gender'*, *'Smoking_Status*', *'Residence'*, *'Air_Pollution_Exposure*', '*Biomass_Fuel_Use*', '*Factory_Exposure*', '*Family_History*', '*Diet_Habit*', '*Symptoms*', '*Histology_Type*', '*Stage*', '*Treatment*', and '*Hospital_Type*'. This crucial preprocessing step transformed each unique category within these features into a new, distinct binary column (either 0 or 1), thereby preventing the machine learning models from incorrectly assuming any inherent ordinal relationships or ranking between the categories. The target variable, *'Survival_1_Year*', which is binary in nature ('Yes' or 'No'), was effectively encoded into a numerical form (specifically, 1 representing 'Yes' and 0 representing 'No')

using the Label Encoding technique provided by scikit-learn.

Train-Test Split

To ensure a robust and unbiased evaluation of the performance of the machine learning models, the preprocessed dataset was systematically split into distinct training and testing sets. A standard and widely accepted split ratio of 70% of the data for training the models and the remaining 30% for independently testing their performance was employed [44]. A fixed *random_state* value of 42 was set during this splitting process. This deliberate choice ensures the complete reproducibility of the data split, thereby allowing for consistent and comparable evaluation of the machine learning models across multiple runs and experiments. This separation guarantees that the models are exclusively trained on a specific subset of the data and subsequently evaluated solely on unseen data, which provides realistic and reliable assessment of their generalization capabilities to new, unobserved patient cases.

Machine Learning Models

Four distinct and widely recognized machine learning models were employed and evaluated in this study with the primary objective of predicting the one-year survival outcome for lung cancer patients. These chosen models represent a diverse range of computational approaches to tackling binary classification problems, spanning from more traditional statistical methods to advanced ensemble techniques and a foundational deep learning architecture.

1. **Logistic Regression:** This model serves as a fundamental linear model specifically designed for binary classification tasks. At its core, it models the probability that a given input data point belongs to a particular class by applying a logistic function to a linear combination of the input features [45]. In this study, Logistic Regression was implemented using the readily available LogisticRegression class, which is part of the sklearn.linear_model module within the scikit-learn library.
2. **Decision Tree:** This model represents a non-linear, tree-like structure that operates by recursively partitioning the input data into increasingly homogeneous subsets based on the values of the input features. Through a series of hierarchical decisions, it constructs a tree structure to arrive at a final class prediction for each data instance [45].
   The DecisionTreeClassifier from the sklearn.tree module in scikit-learn was utilized for its implementation in this research.
3. **Random Forest:** As a powerful ensemble learning method, Random Forest operates by constructing a large multitude of individual decision trees during the training phase. For classification tasks, it aggregates the predictions from all the individual trees (typically by taking a majority vote or determining the mode of the predicted classes) to output the final class prediction [46]. This ensemble approach is well-known for its effectiveness in reducing overfitting and improving robustness compared to relying on a single decision tree.
   The RandomForestClassifier from the sklearn.ensemble module was employed for this model.
4. **Gradient Boosting:** This is another highly effective ensemble technique. Unlike Random Forests which build trees independently, Gradient Boosting builds trees sequentially and iteratively. Each new tree in the ensemble is specifically trained to correct the errors or misclassifications

made by the combined predictions of all previously built trees. It effectively combines the predictions of multiple "weak learners" (which are typically shallow decision trees) to create a single, powerful "strong learner" [47]. The XGBClassifier from the xgboost library, which is an optimized and highly efficient implementation of gradient boosting, was used in this study.

5. **Neural Network:** A basic, yet illustrative, feedforward neural network model was constructed to explore the capabilities of deep learning for this prediction task. This type of model consists of interconnected layers of artificial nodes or "neurons". These neurons process the input features through a series of weighted connections and non-linear activation functions to learn intricate, complex patterns and ultimately make predictions [45]. A sequential model architecture comprising multiple dense layers was implemented using the high-level Keras API, which is integrated within TensorFlow. The model design included an input layer, one or more hidden layers with the 'relu' activation function (Rectified Linear Unit) for introducing non-linearity, and a final output layer with a 'sigmoid' activation function specifically suitable for binary classification tasks, as it outputs a probability between 0 and 1.

RESULTS AND DISCUSSION

Data Exploration

Exploratory data analysis was conducted to gain a comprehensive understanding of the distributions of key features within the dataset and to examine their relationships with the survival outcome variable. This initial phase is crucial for identifying patterns, understanding data characteristics, and informing subsequent modeling steps.

The distribution of patient age revealed a varied age range within the dataset, encompassing a broad spectrum of adult age groups commonly affected by lung cancer.

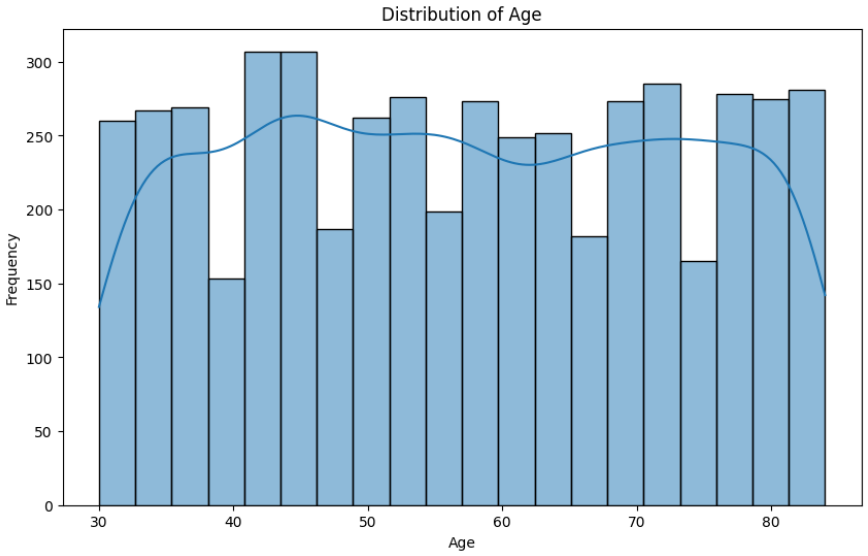


Fig. 1: Distribution of Patient Age from Dataset

A comparative analysis of the age distribution between the two survival groups ('Yes' for survival and 'No' for non-survival) indicated remarkably similar age profiles between those who survived and those who did not. This initial observation suggests that age, when considered in isolation, might not be a particularly strong independent predictor of one-year survival within the context of this specific dataset.

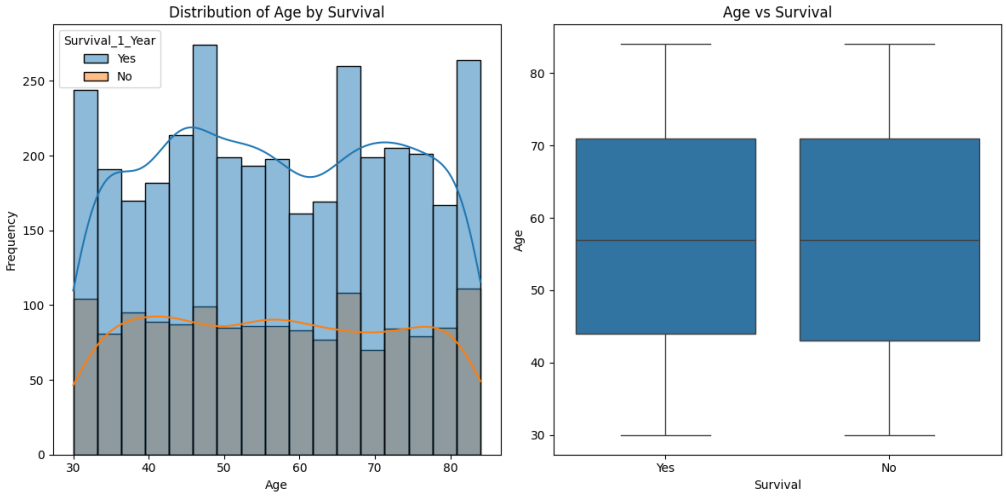


Fig 2: Comparative Analysis of the Age Distribution between the Two Survival Groups

Analysis of the gender distribution within the dataset clearly showed a higher number of male patients compared to female patients, reflecting potential demographic patterns or prevalence rates in the studied population.

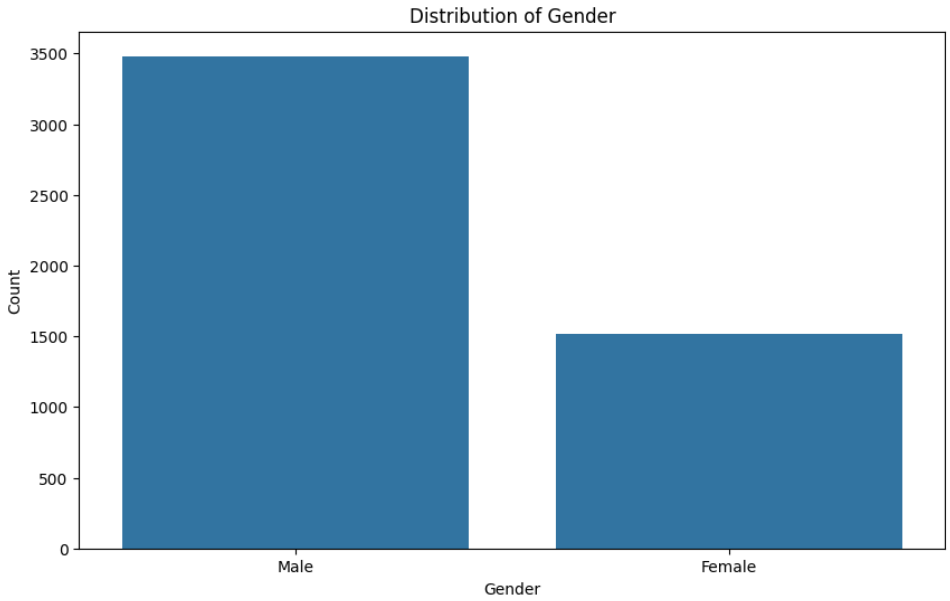


Fig 3: Gender Distribution of Patients within the Dataset

Examining the relationship between gender and survival through a stratified count plot revealed potential differences in one-year survival rates between males and females. This finding suggests that gender might indeed play a role in influencing patient outcomes, possibly due to biological differences, varying exposure patterns, or disparities in healthcare access.

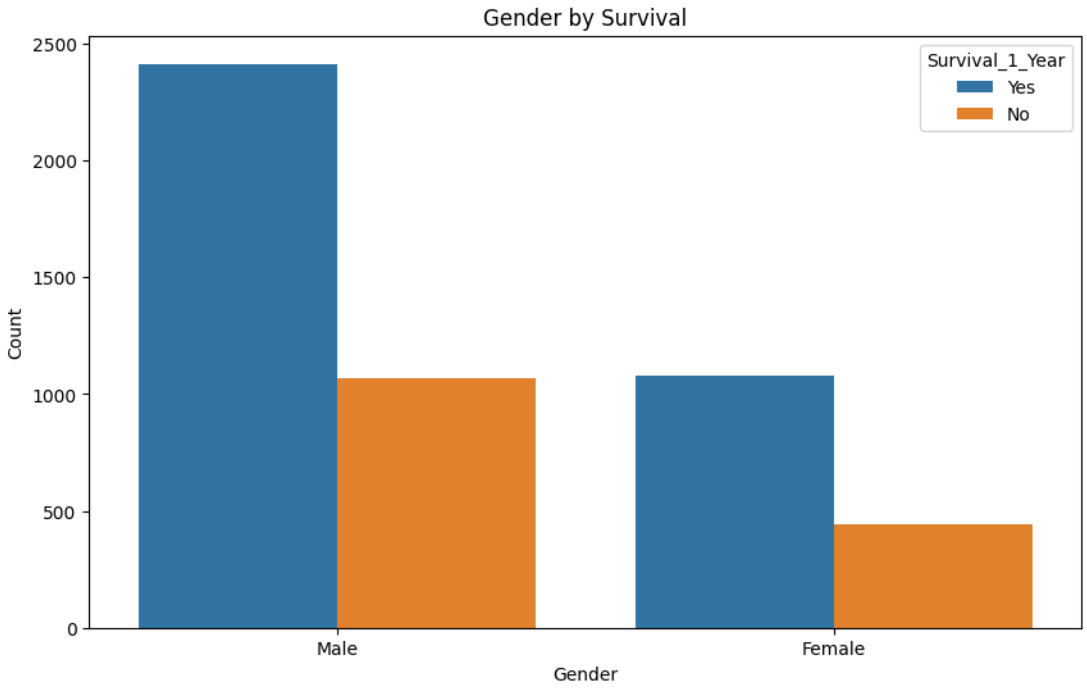


Fig 4: Relationship between Gender and Survival

The distribution of smoking status among the patients indicated that 'Current Smokers' constituted the largest and most prevalent group in the dataset, underscoring the significance of this risk factor in the studied population.

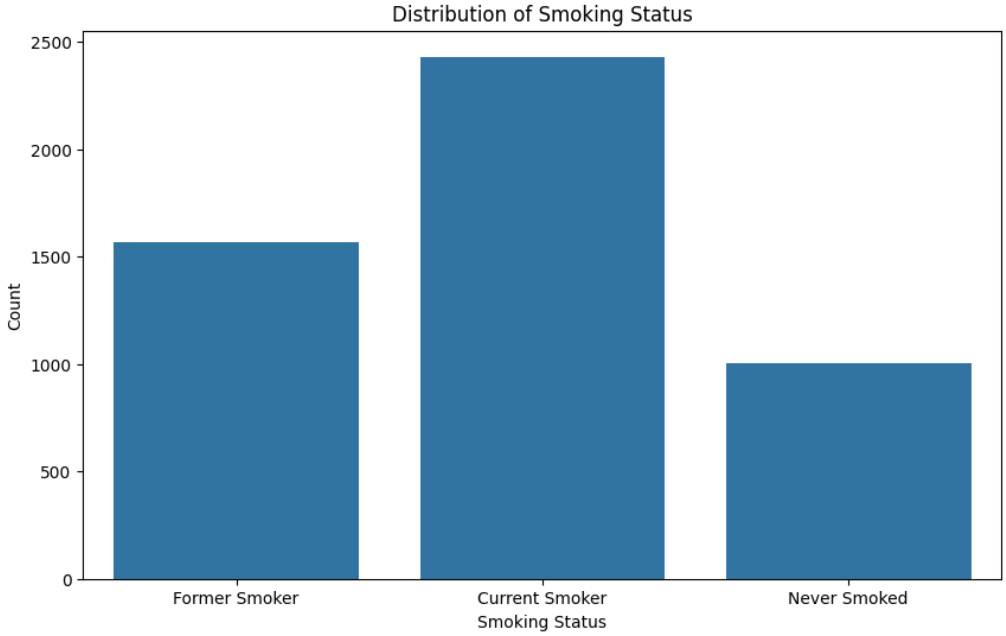


Fig 5: Distribution of Smoking Status of Patients from the Data

Survival analysis stratified by smoking status demonstrated varying survival rates across the 'Former Smoker', 'Current Smoker', and 'Never Smoked' categories. This finding highlights the well-established impact of smoking on lung cancer prognosis, where continued or past smoking history generally correlates with less favorable outcomes.

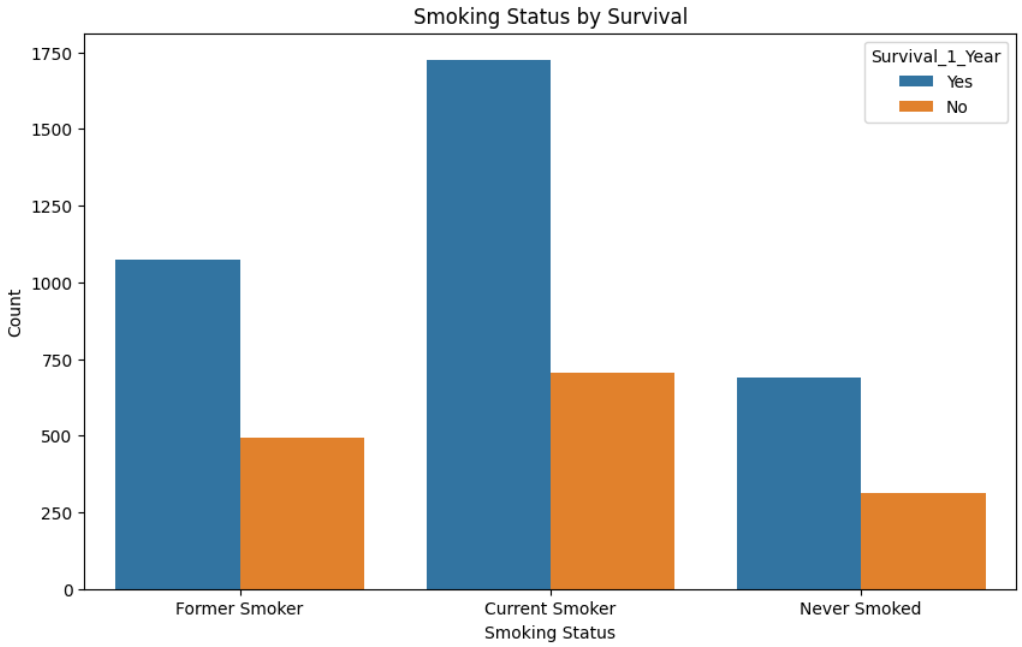


Fig 6: Survival Analysis by Smoking Status of Patients

The distribution of patient residence showed that a larger proportion of patients were from urban areas compared to rural areas in this dataset.

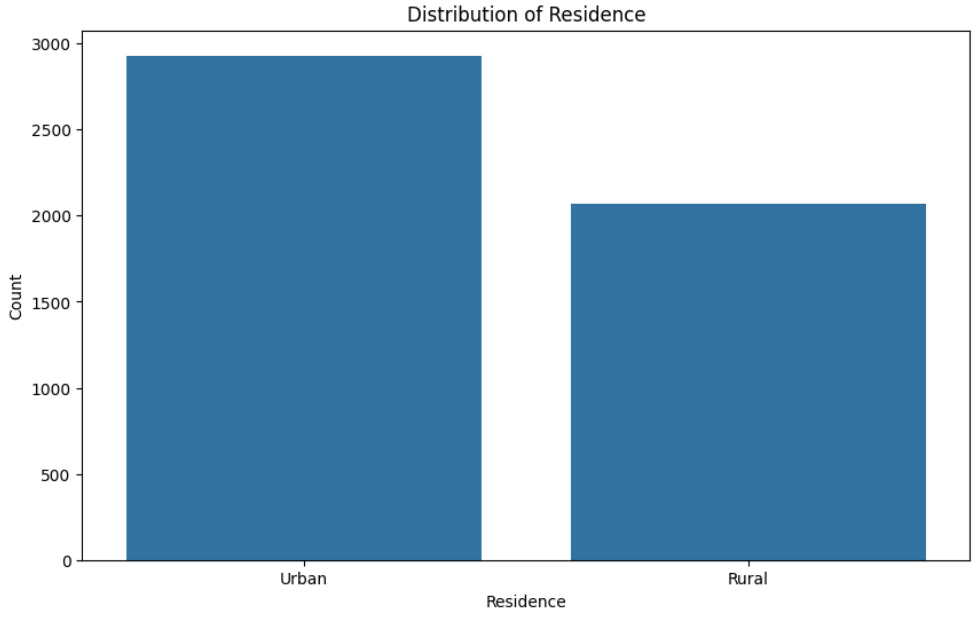


Fig 7: Distribution of Patient Residence

Survival analysis examining residence suggested potential differences in survival rates between urban and rural residents. These differences could be attributed to variations in environmental exposures, lifestyle factors, or access to specialized healthcare facilities.

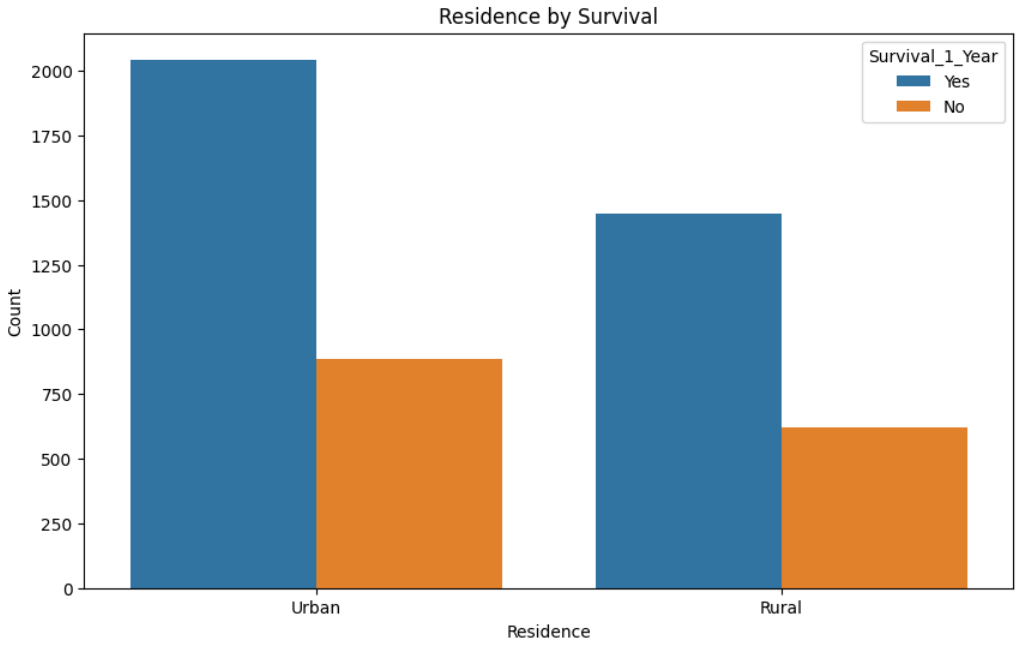


Fig 8: Survival analysis by Residence of Patients

Air pollution exposure was categorized into three levels: 'Low', 'Moderate', and 'High'. The distribution showed that 'High' exposure was the most commonly reported level among the patients.

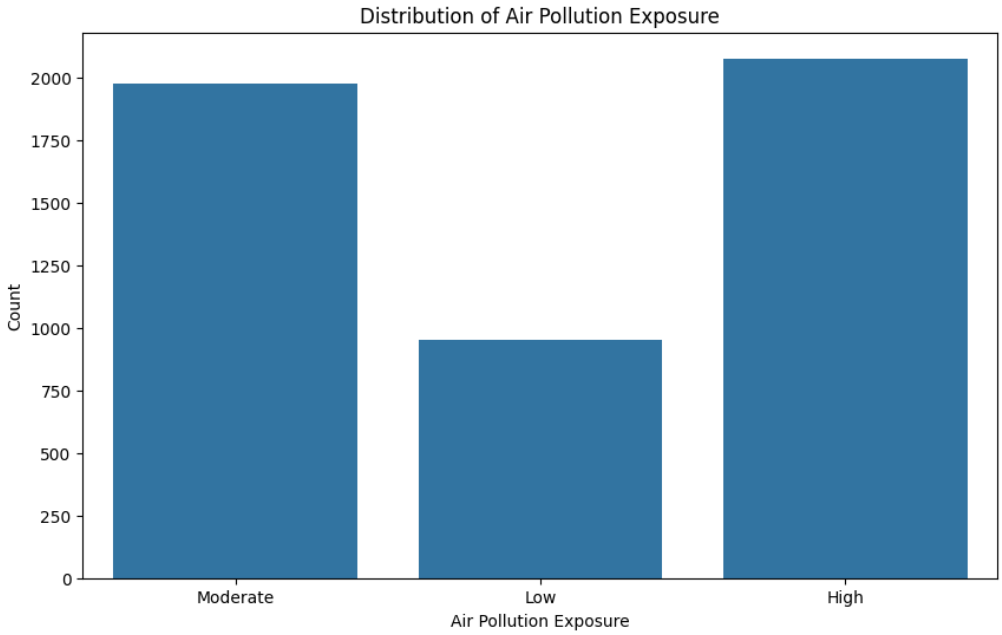


Fig 9: Air Pollution Exposure Distribution

A notable relationship was observed between air pollution exposure levels and residence, with urban areas clearly showing a higher prevalence of moderate and high air pollution exposure compared to rural areas. This correlation aligns with typical patterns of environmental pollution in urbanized settings.

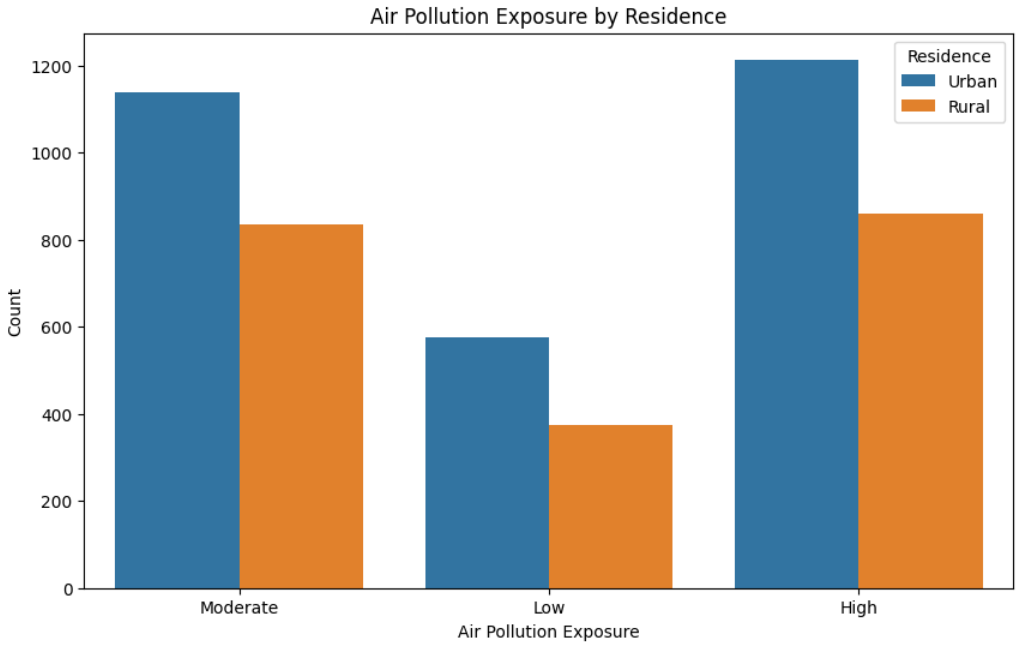


Fig 10: Air Pollution Exposure by Residence

Survival analysis stratified by air pollution exposure indicated potential differences in survival rates across the different exposure levels. Higher exposure to air pollution is a known risk factor and its impact on survival is an important consideration.

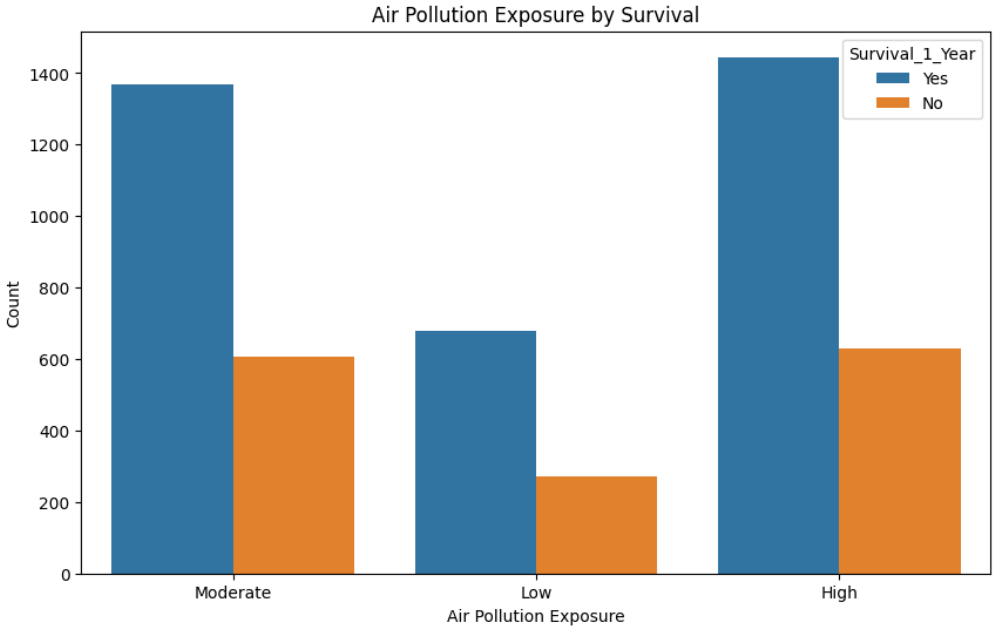


Fig 11: Air Pollution Exposure by Survival

Biomass fuel use was reported by a significant portion of the dataset's patient population. This factor is often associated with indoor air pollution, particularly in certain residential settings.

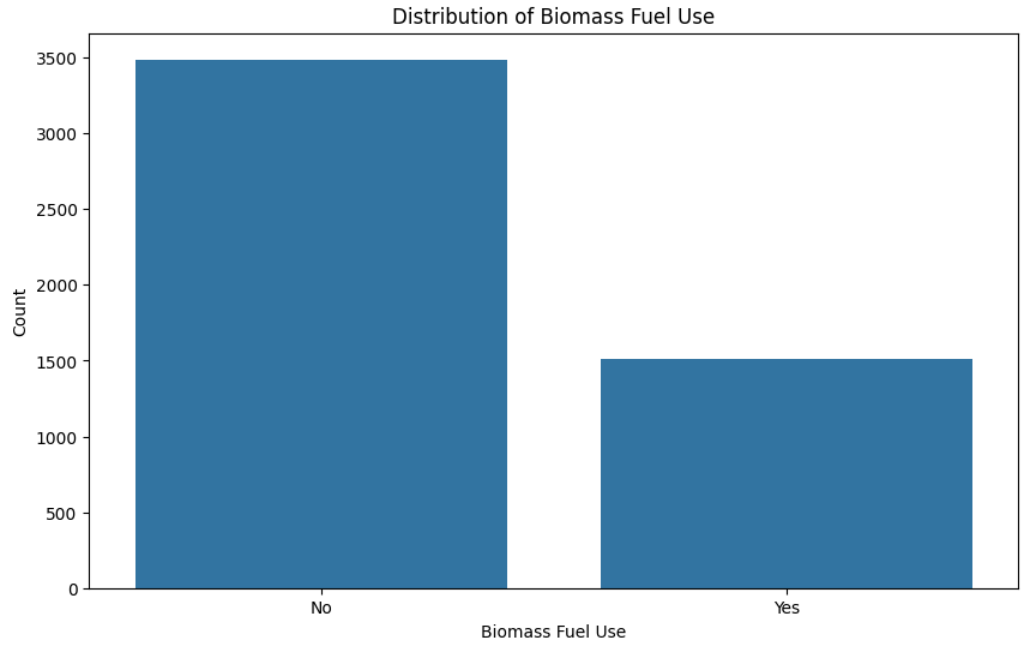


Fig 12: Distribution of Biomass Fuel Use

Biomass fuel use was found to be strongly correlated with rural residence, as expected, highlighting a specific environmental exposure concentrated in non-urban areas.

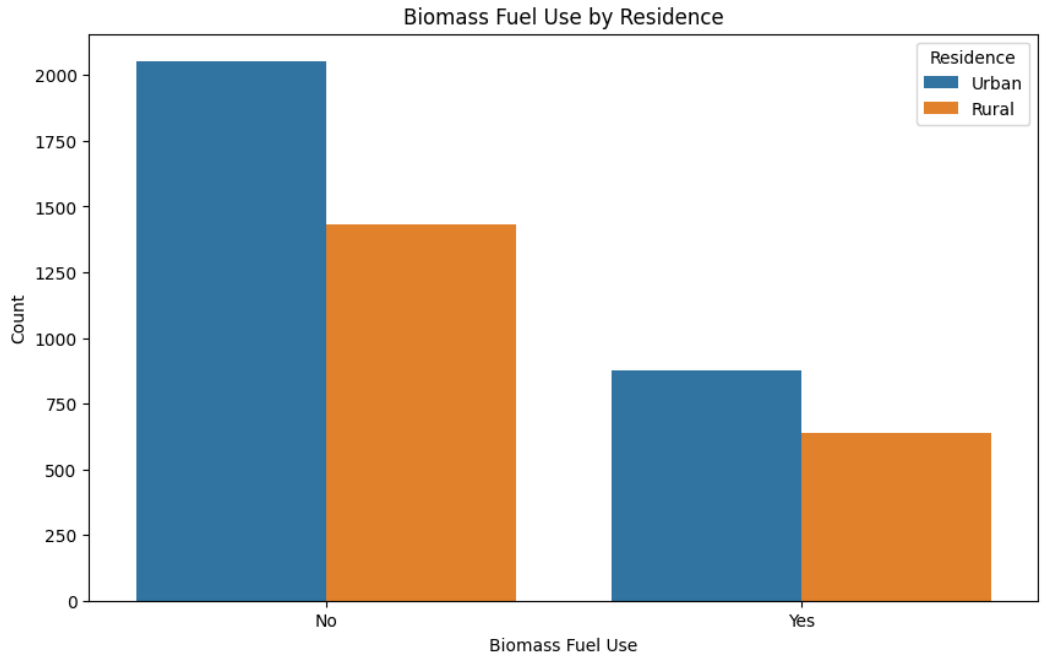


Fig 13: Biomass Fuel Use by Residence

Survival analysis examining biomass fuel use suggested potential differences in survival outcomes between individuals who reported using biomass fuel and those who did not, indicating a possible link between this exposure and prognosis.

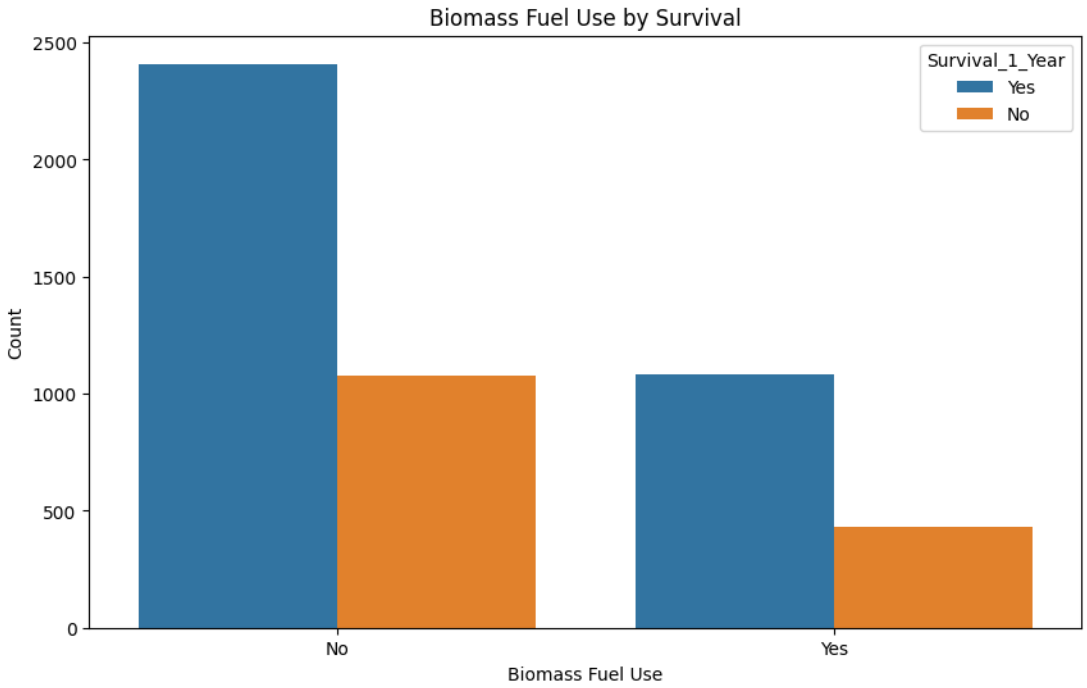


Fig 14: Biomass Fuel Use by Survival

Factory exposure, representing potential occupational or environmental exposure to industrial pollutants, was reported by a smaller portion of the dataset's population compared to other risk factors.

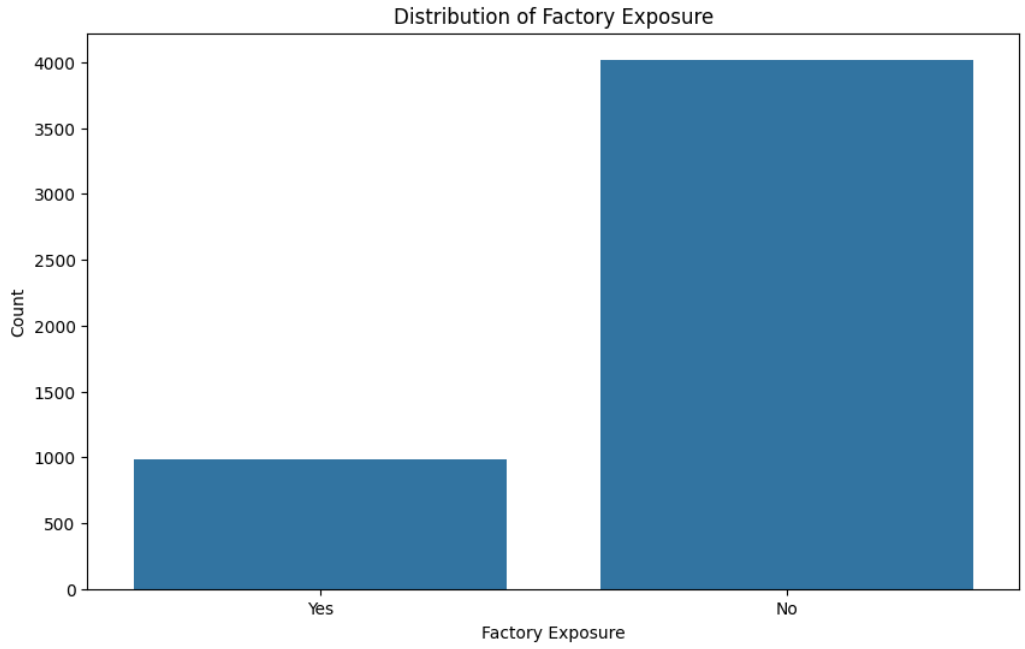


Fig 15: Distribution of Factory Exposure

Survival analysis stratified by factory exposure indicated potential differences in survival outcomes between individuals with reported factory exposure and those without. This suggests that such industrial exposures may influence prognosis.

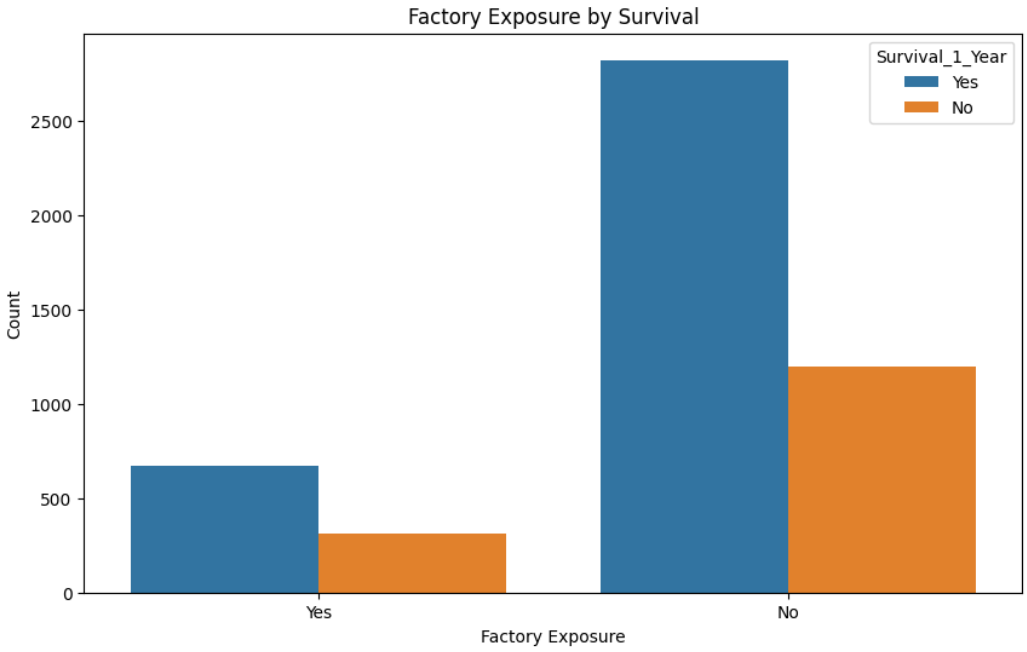


Fig 16: Factory Exposure by Survival

A family history of lung cancer was reported by a minority of patients in the dataset. While not as prevalent as factors like smoking, family history is a significant indicator of potential genetic predisposition.

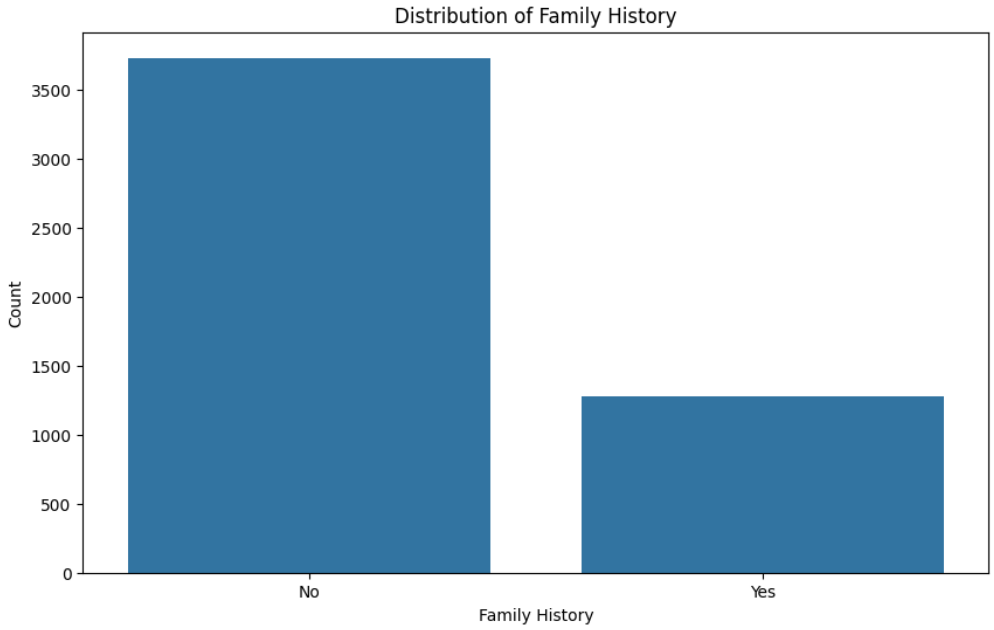


Fig 17: Distribution of Family History

Survival analysis examining family history suggested potential differences in survival outcomes for individuals with a reported family history of lung cancer. This finding supports the idea that inherited factors might influence prognosis.

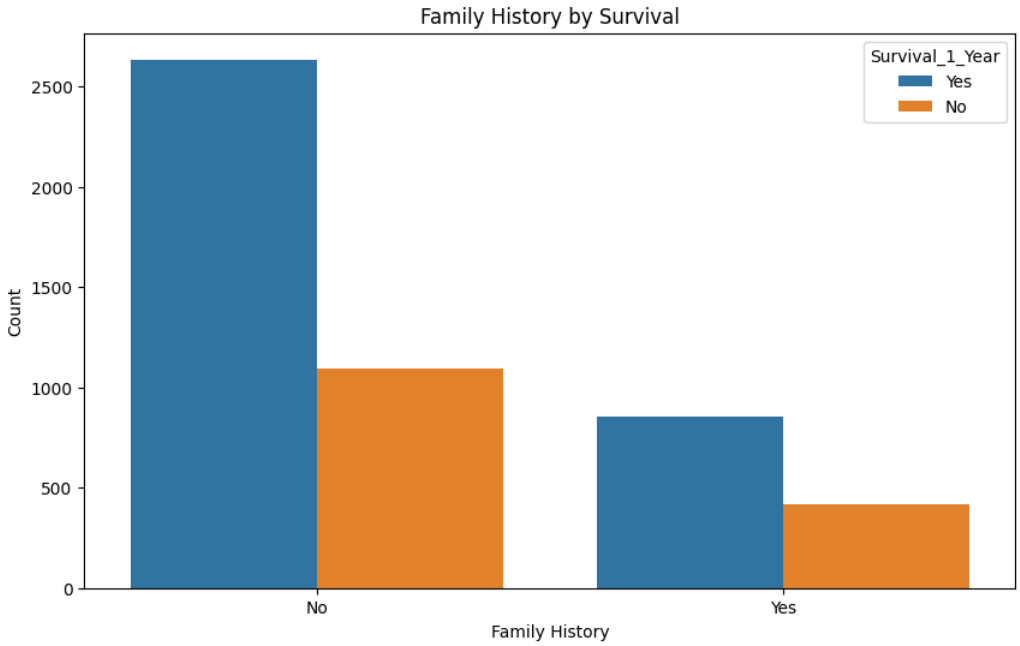


Fig 18: Family History by Survival

The distribution of diet habits among the patients showed 'Non-Vegetarian' as the largest category, followed by 'Mixed' and 'Vegetarian'.

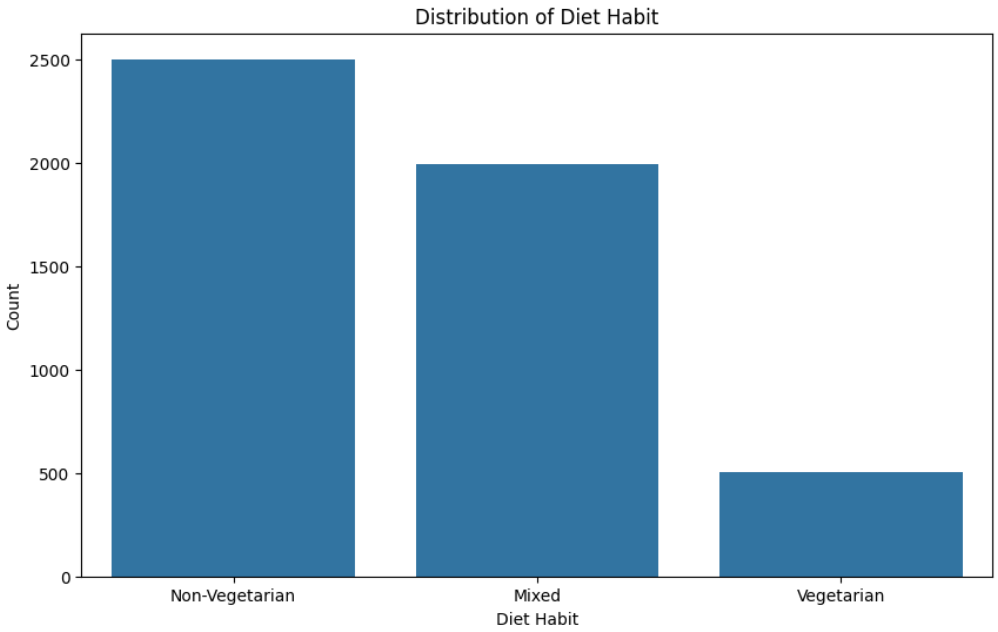


Fig 19: Distribution of Diet Habit

Survival analysis stratified by diet habit indicated potential differences in survival rates across the different dietary groups, suggesting that nutritional factors may play a role in patient outcomes.

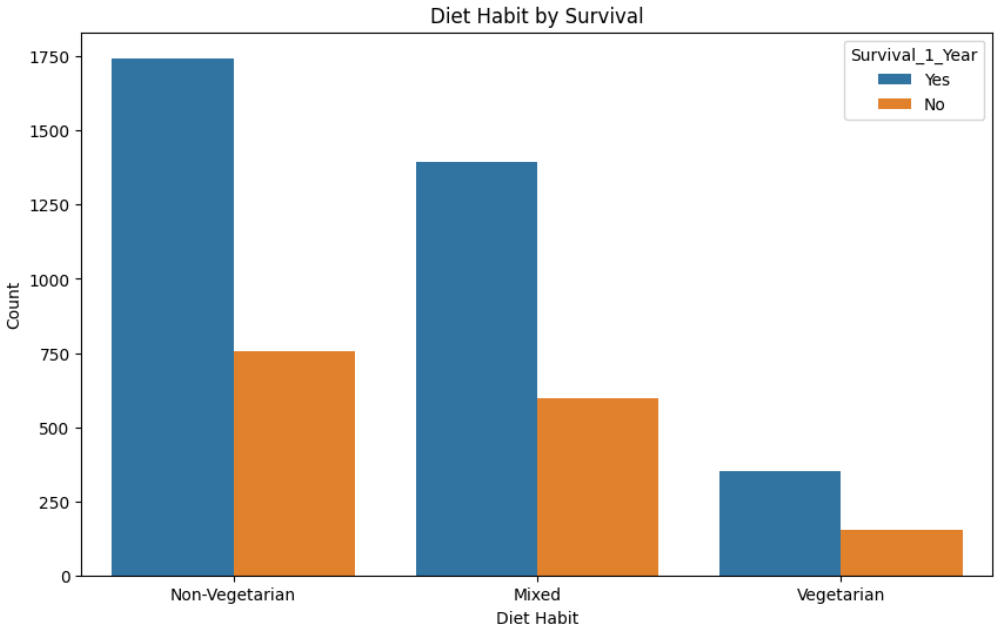
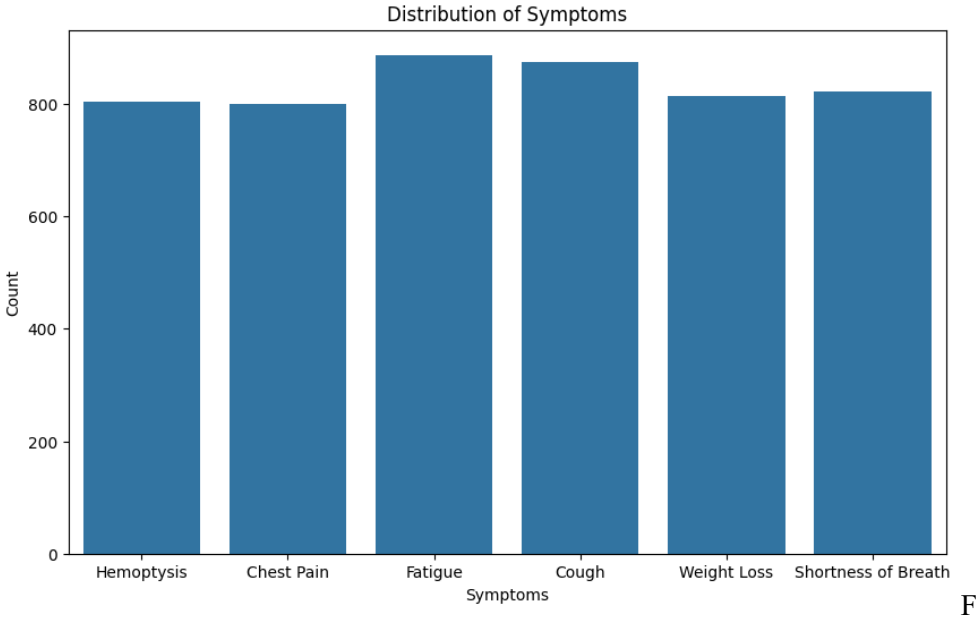


Fig 20: Diet Habit by Survival

The distributions of reported symptoms at the time of diagnosis were relatively balanced across the categories, including Hemoptysis, Chest Pain, Fatigue, Cough, Weight Loss, and Shortness of Breath.



F

ig 21: Distribution of Symptoms

Survival analysis examining symptoms suggested potential differences in survival based on the specific presenting symptoms, as certain symptoms may be indicative of more advanced or aggressive disease at diagnosis.

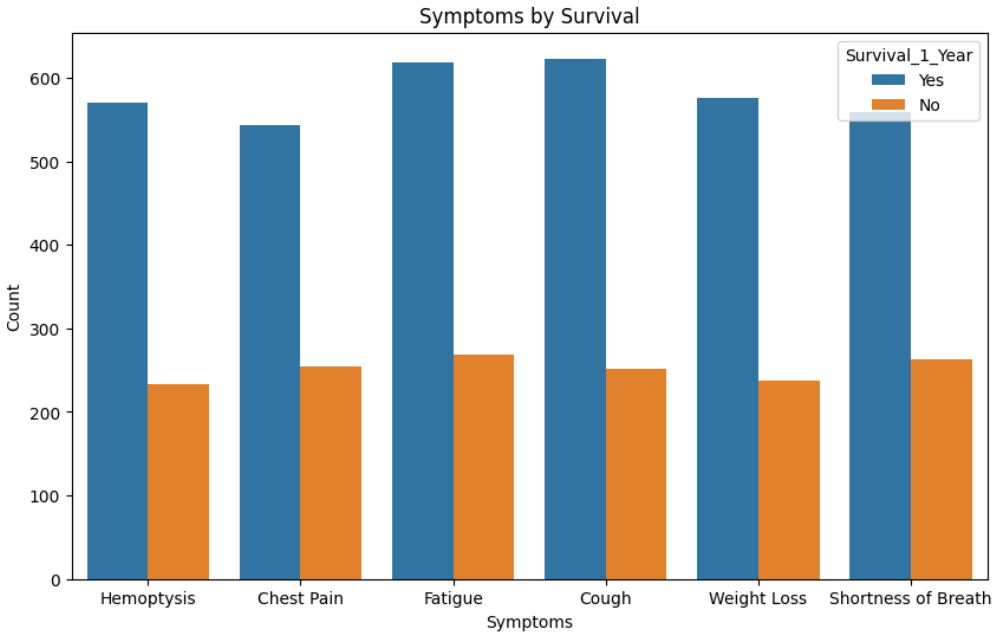


Fig 22: Symptoms by Survival

The distribution of tumor size in millimeters showed a spread of tumor sizes among the patient cohort, reflecting the variability in disease presentation.

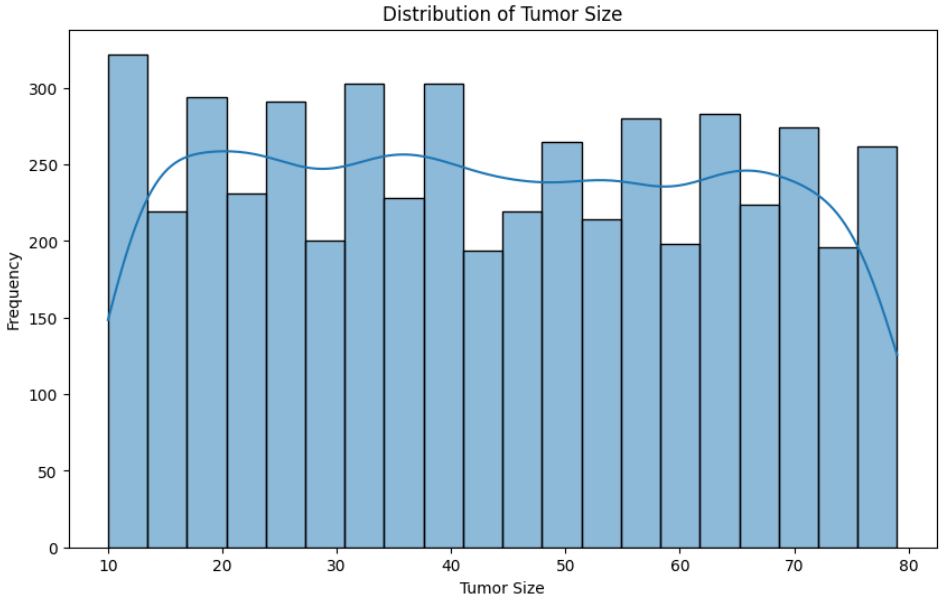


Fig 23: Distribution of Tumor Size

Survival analysis examining the distribution of tumor size stratified by survival outcome suggested discernible differences in survival based on the tumor size at diagnosis. Generally, smaller tumor sizes are clinically associated with better prognoses, and this pattern appeared to be reflected in the data.

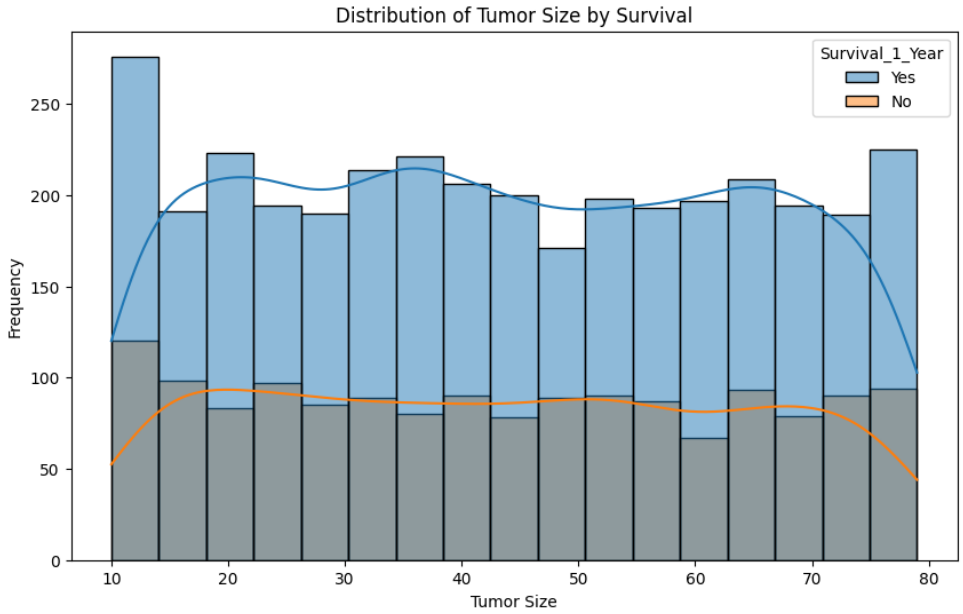


Fig 24: Distribution of Tumor Size by Survival

The histology type distribution clearly showed 'Adenocarcinoma' as the most common type of lung cancer within this dataset, consistent with global epidemiological trends.

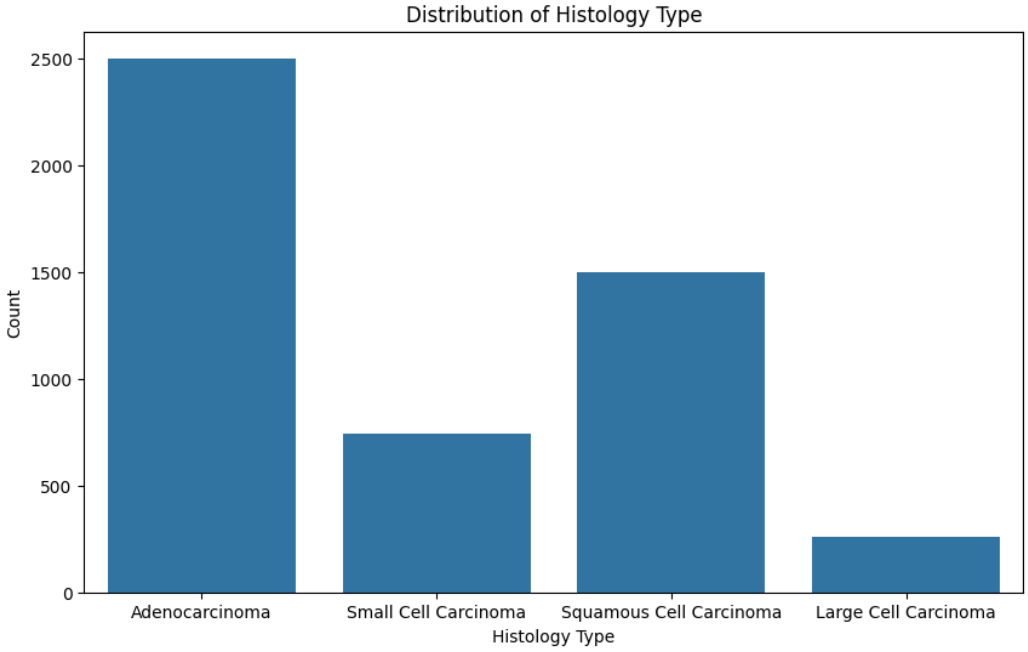


Fig 25: Distribution of Histology Type

Survival analysis stratified by histology type indicated significant differences in one-year survival rates across the different histology types. This finding is expected, as different histology types exhibit varying biological behaviors and responsiveness to treatment.

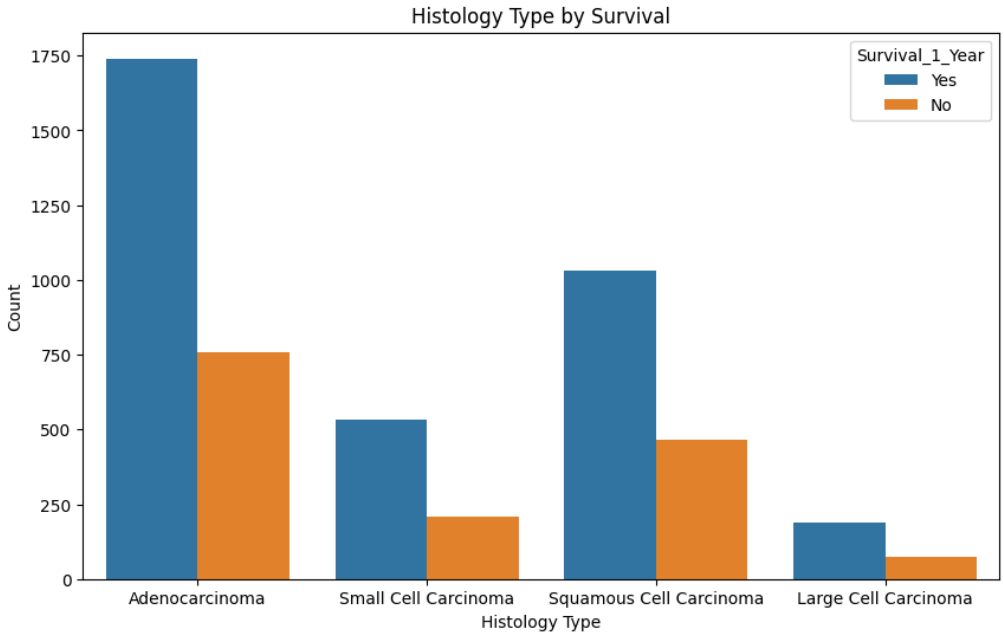


Fig 26: Histology Type by Survival

The cancer stage distribution at diagnosis showed higher patient counts for Stages II and III, with fewer patients diagnosed at Stage I and Stage IV.

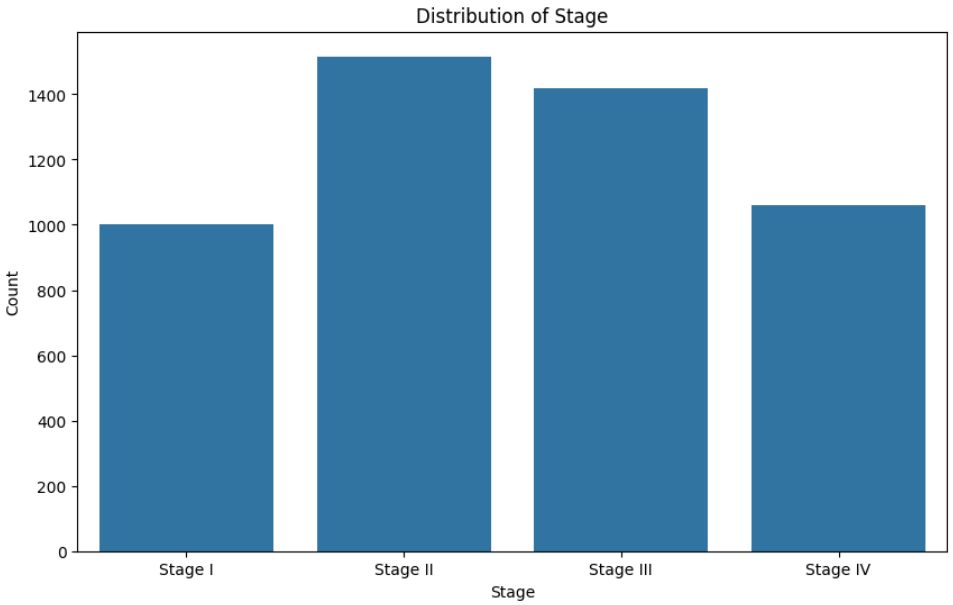


Fig 27: Distribution of Stage

Survival analysis by stage at diagnosis clearly and strongly showed that patients diagnosed at earlier stages (Stage I and Stage II) had a significantly higher proportion of survivors compared to those diagnosed at later stages (Stage III and Stage IV). This finding unequivocally confirms cancer stage as a powerful and critical prognostic factor.

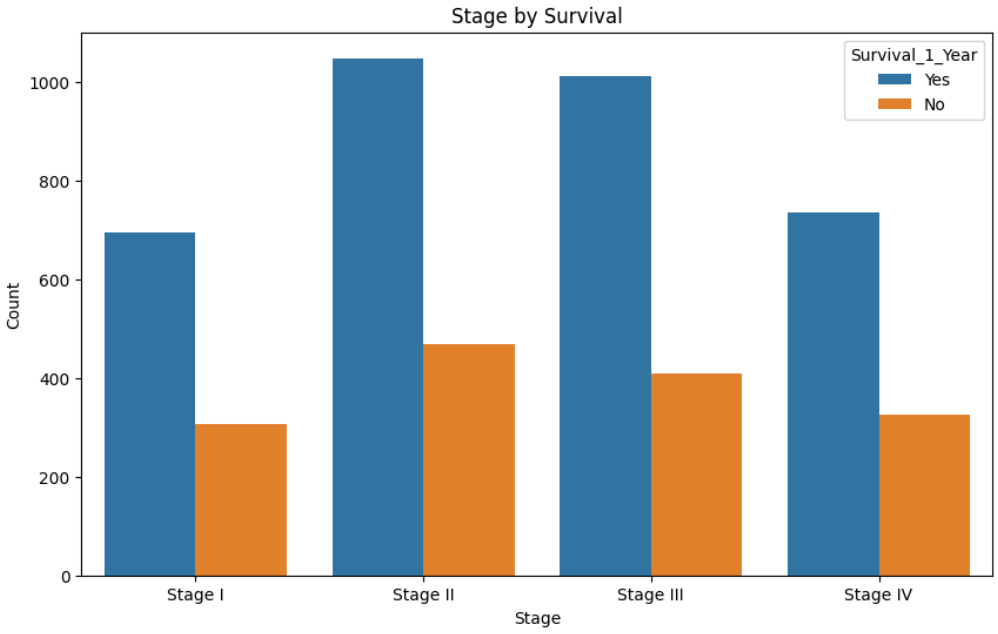


Fig 28: Stage by Survival

The distribution of treatments received by the patients showed 'Chemotherapy' as the most frequently administered treatment modality in this dataset.

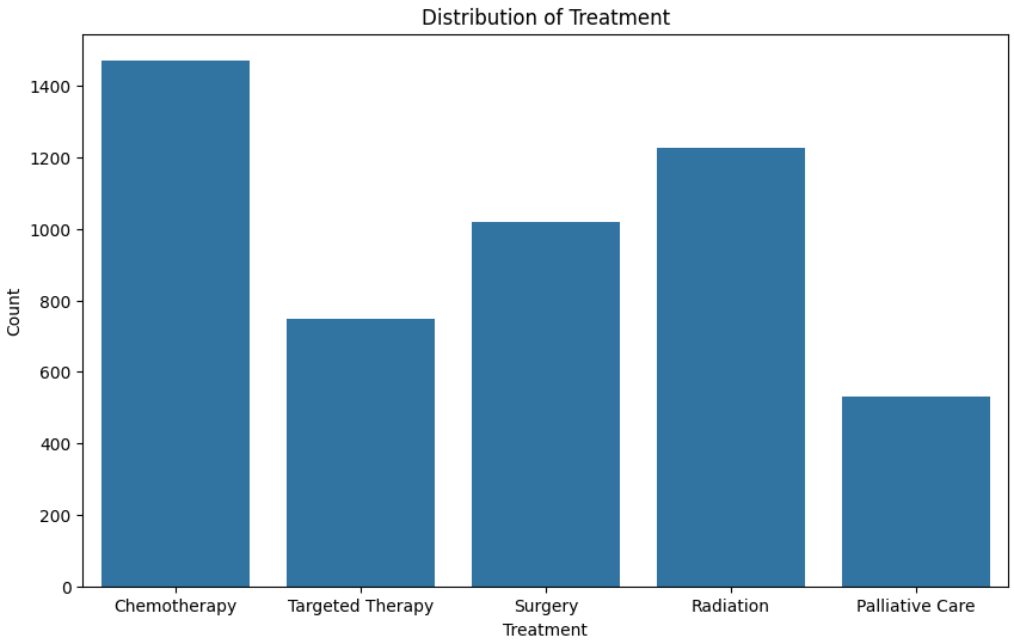


Fig 29: Distribution of Treatment

Survival analysis stratified by the type of treatment indicated differences in survival rates based on the specific treatment received. This is expected, as treatment effectiveness varies depending on the stage, histology, and other patient characteristics.

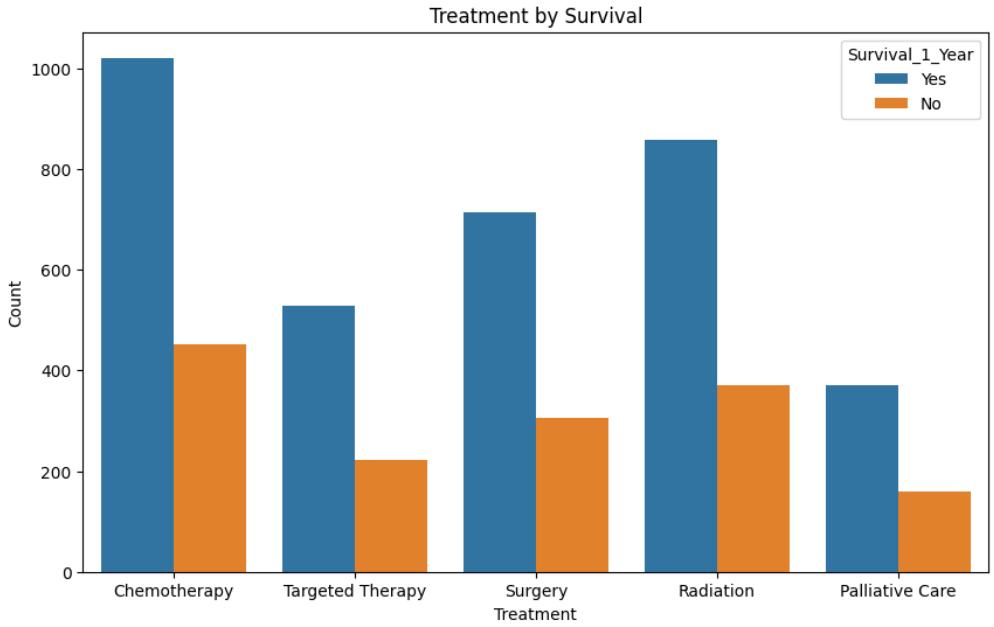


Fig 30: Treatment by Survival

The distribution of hospital types where patients received treatment showed that 'Government' hospitals treated the highest number of patients in this dataset.

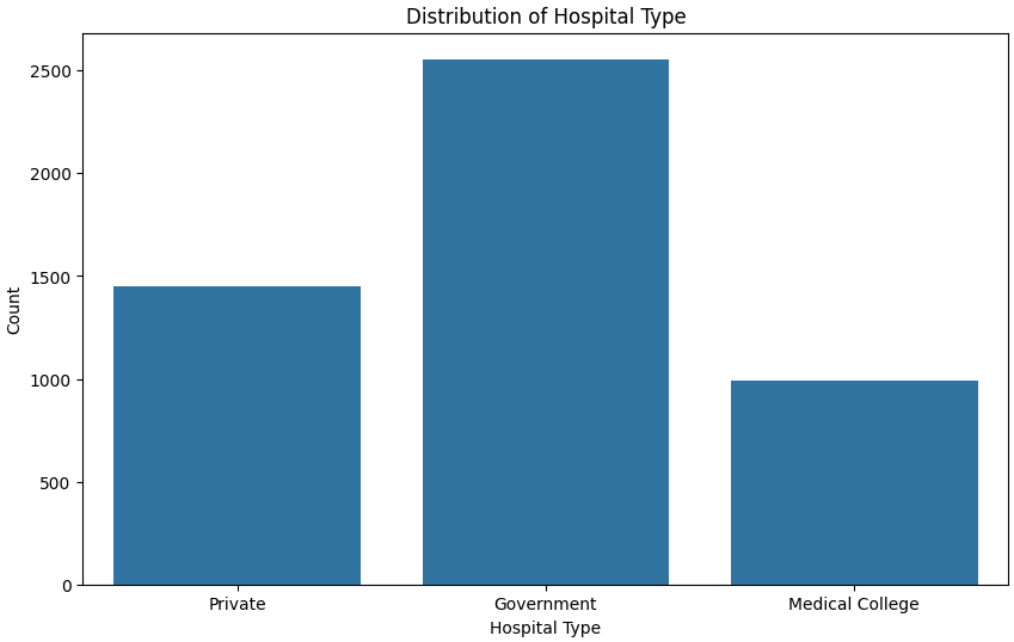


Fig 31: Distribution of Hospital Type

A relationship was observed between hospital type and patient residence, with urban areas potentially having greater access to certain types of healthcare facilities. This highlights potential geographical disparities in healthcare access.

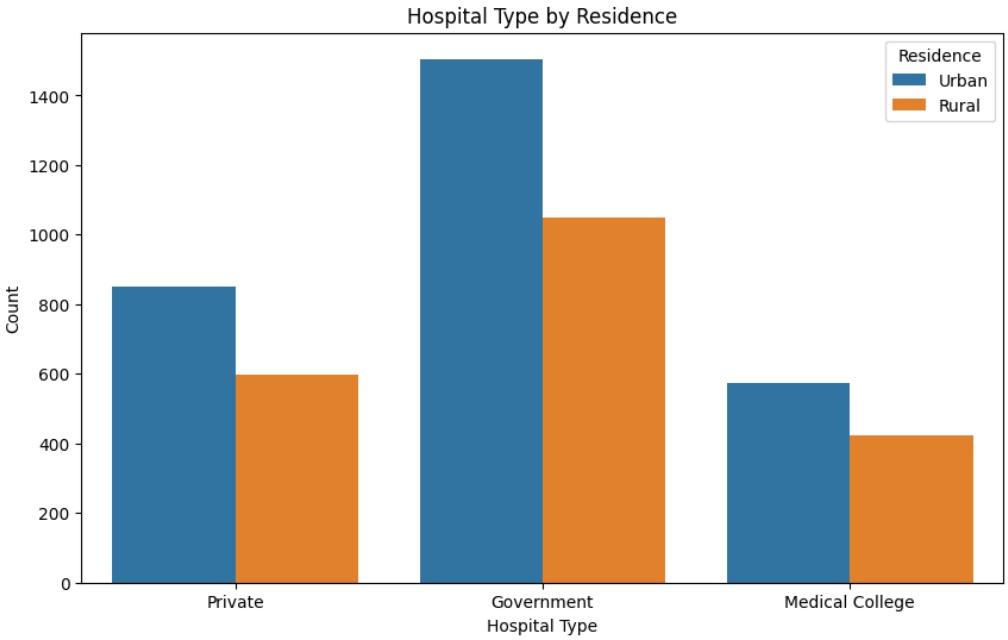


Fig 32: Hospital Type by Residence

Survival analysis examining hospital type suggested potential differences in survival outcomes across different healthcare settings. These differences could be influenced by variations in available resources, specialized expertise, and treatment protocols.

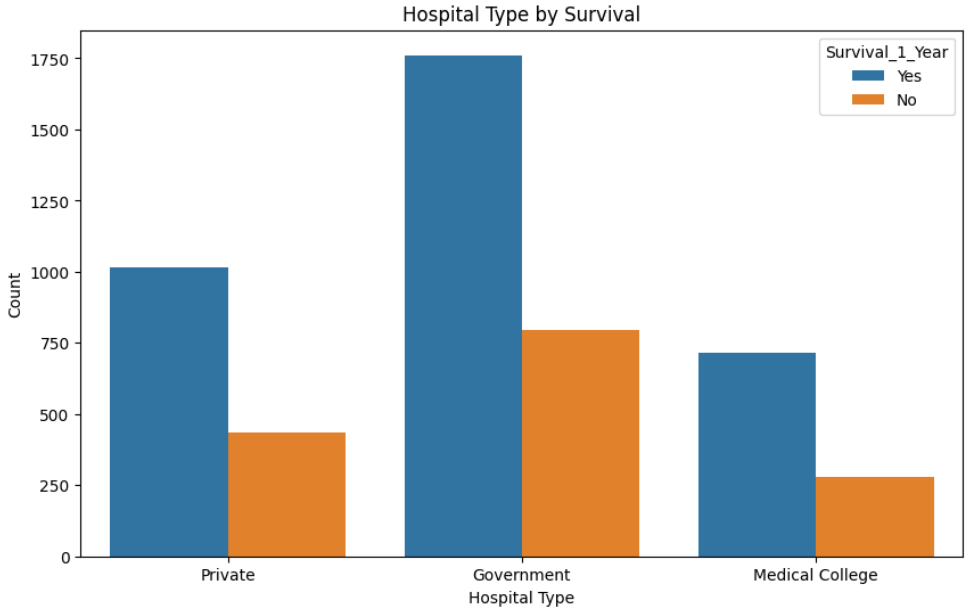


Fig 33: Hospital Type by Survival

Survival Probability Estimation

Beyond simple binary prediction, estimating the probability of survival provides a more nuanced understanding of risk [48]. The predicted survival probabilities from the Logistic Regression, Gradient Boosting, and Neural Network models were analyzed to understand the distribution of risk estimated by each model across the test set.

The Logistic Regression model's predicted survival probabilities showed a relatively concentrated distribution, primarily centered

around a probability of approximately 0.7. This suggests that the model is predicting a moderate to high likelihood of survival for a large portion of the test instances.

In contrast, the Gradient Boosting model's predicted survival probabilities exhibited a distribution noticeably skewed towards higher probabilities, with a prominent peak near 0.9. This pattern suggests that the Gradient Boosting model is more confident in predicting a higher likelihood of survival for a larger proportion of the test set compared to the Logistic Regression model.

The Neural Network model's predicted survival probabilities showed a distribution somewhat similar in shape to that of the Logistic Regression model, with a central tendency, but perhaps slightly skewed towards higher probabilities. This indicates that the Neural Network's probability estimations are also somewhat concentrated, although potentially leaning more towards positive survival outcomes.

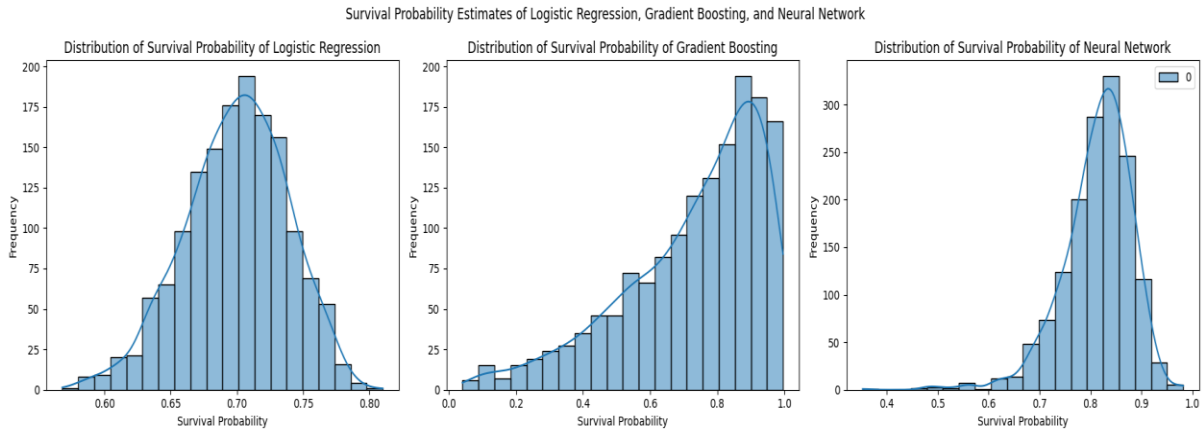


Fig 34: Survival Probability Estimates of Logistic Regression, Gradient Boosting, and Neural Network

The observed differences in the distributions of predicted probabilities across these models highlight that the models are not only capturing different patterns within the data but also expressing varying levels of certainty and bias in their risk estimations. This underscores the importance of examining probability outputs in addition to binary predictions for a complete understanding of model behavior.

Survival Prediction Model Performance

Four distinct machine learning models were rigorously evaluated for their capability to accurately predict one-year survival outcomes. The performance of each model was assessed using standard classification metrics, and confusion matrices were generated to provide a detailed breakdown of their predictions.

Model	Accuracy	F1 Score	Precision	Recall
Decision Tree	0.565	0.680	0.696	0.666
Random Forest	0.684	0.811	0.695	0.974
Gradient Boosting	0.641	0.768	0.698	0.852
Neural Network	0.681	0.808	0.695	0.965

Table 1: Performance Metrics Analysis

The calculated performance metrics provided a clear basis for comparing the models. The Random Forest and Neural Network models consistently achieved the highest accuracy and F1 scores among the evaluated models. This indicates their superior overall capability in

making correct survival predictions, balancing both precision and recall. All four models demonstrated relatively similar precision scores, suggesting that when they predict a patient will survive ('Yes'), they are correct a comparable proportion of the time. However, the Random Forest and Neural Network models distinguished themselves by demonstrating significantly higher recall scores. This high recall signifies that these models were particularly effective at identifying the vast majority of patients who actually survived in the test set. The Gradient Boosting model showed moderate performance, while the Decision Tree consistently exhibited the lowest performance across most metrics.

A detailed examination of the confusion matrices provided deeper insights into the types of errors made by each model:

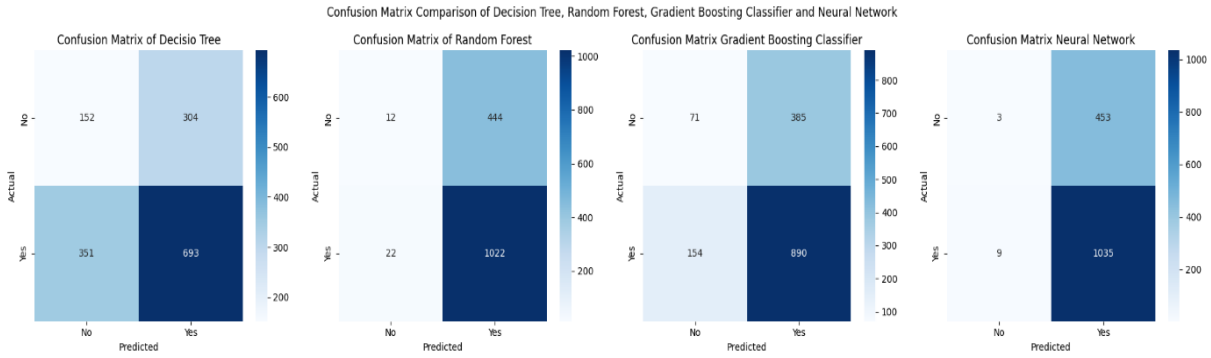


Fig 35: Confusion Matrix Comparison of Decision Tree, Random Forest, Gradient Boosting Classifier and Neural Network

The confusion matrix for the Decision Tree confirmed its lower recall, showing a substantial number of False Negatives (instances where the model incorrectly predicted 'No' survival for patients who actually survived). The confusion matrices for the Random Forest and Neural Network models highlighted their strength in minimizing False Negatives, aligning with their high recall. However, they also revealed a notable trade-off: a relatively higher number of False Positives (instances where the model incorrectly predicted 'Yes' survival for patients who did not survive). The Gradient Boosting model's confusion matrix indicated a more balanced performance, with fewer False Positives than the Random Forest and Neural Network, and fewer False Negatives than the Decision Tree, suggesting a better compromise in correctly identifying both survivors and non-survivors.

Feature Importance Analysis

Understanding which features are most influential in the prediction process is crucial for gaining insights into the underlying factors driving survival outcomes and for informing precision prevention strategies. Feature importance analysis was conducted for the tree-based models: Decision Tree, Random Forest, and Gradient Boosting. (Note: Standard Neural Networks do not provide readily interpretable feature importances in the same direct manner as these models.)

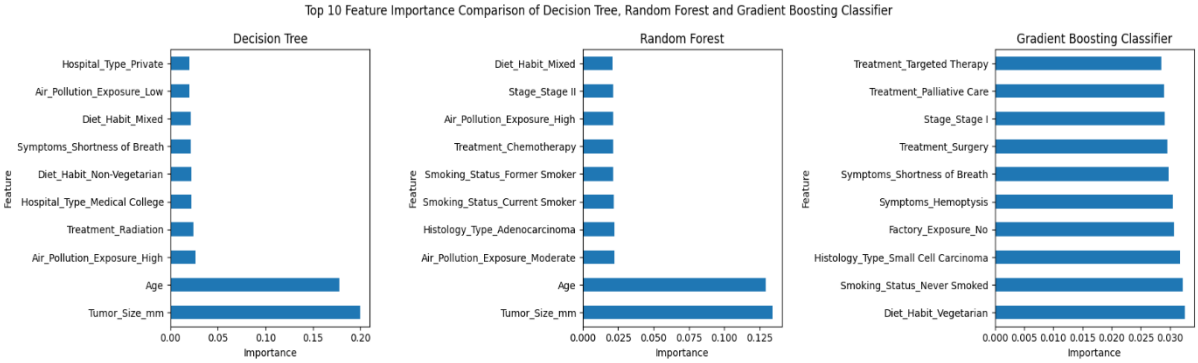


Fig 36: Top 10 Feature Importance Comparison of Decision Tree, Random Forest and Gradient Boosting Classifier

For both the Decision Tree and Random Forest models, traditional clinical factors such as Tumor_Size_mm and Age consistently emerged as the top most important features influencing survival prediction. This aligns well with established medical knowledge regarding the strong prognostic value of tumor characteristics and patient age. Beyond these, other features appearing in the top importance lists for these models included various lifestyle factors (e.g., smoking status, diet habit), environmental exposures (air pollution, factory exposure), specific symptoms reported (e.g., cough), histology type, cancer stage, the type of treatment received (e.g., chemotherapy), and the hospital type. The specific ranking and the exact set of other features in the top importance list showed some variation between the Decision Tree and Random Forest models, reflecting the differences in how these algorithms weight feature contributions.

The Gradient Boosting model, however, presented a distinct and particularly interesting perspective on feature importance. In this model, lifestyle factors like *Diet_Habit_Vegetarian* and *Smoking_Status_Never* Smoked were ranked highly among the most important features. Other influential factors in the Gradient Boosting model's top list included specific histology types (*Histology_Type_Small* Cell Carcinoma), environmental factors (*Factory_Exposure_No*), specific symptoms (*Symptoms_Hemoptysis*, *Symptoms_Shortness* of Breath), treatment types (*Treatment_Surgery*, *Treatment_Palliative* Care, *Treatment_Targeted* Therapy), and cancer stage (*Stage_Stage I*). Notably, *Tumor_Size_mm* and *Age*, which were the dominant features in the Decision Tree and Random Forest models, did not appear in the top 10 most important features for the Gradient Boosting model. This significant difference in feature importance rankings across the models underscores that different machine learning algorithms may capture and utilize distinct relationships and contributions of various features to the prediction task, potentially highlighting more complex or interactive effects in the case of Gradient Boosting.

DISCUSSION OF FINDINGS AND IMPLICATIONS

The exploratory data analysis phase of this study successfully confirmed the relevance of a variety of clinical, demographic, and environmental factors to both the presence of lung cancer and, crucially, patient survival outcomes. These findings largely align with established medical knowledge and epidemiological observations regarding the key determinants of lung cancer prognosis [49]. The analysis of survival probability

distributions from the different machine learning models provided varying perspectives on the estimated risk landscape of the patient cohort, highlighting differences in model confidence and prediction profiles.

The performance evaluation of the survival prediction models indicated that ensemble methods (Random Forest, Gradient Boosting) and the Neural Network generally demonstrated superior predictive capabilities compared to the simpler single Decision Tree model. The particularly high recall achieved by the Random Forest and Neural Network models is a valuable characteristic in a clinical context, as it signifies their strong ability to identify the vast majority of actual survivors. This can be particularly useful for assessing the potential positive impact of interventions or for identifying patient subgroups with favorable prognoses. However, the tendency of these models for a higher rate of false positives, incorrectly predicting survival for patients who do not survive, is a critical consideration. Such misclassifications can have significant clinical implications, potentially leading to less aggressive treatment or inadequate supportive care. The Gradient Boosting model, while having slightly lower recall than the top two, offered a better balance between minimizing false positives and false negatives, representing a more clinically practical trade-off for balanced prediction.

The feature importance analysis conducted across the different models provided crucial and sometimes divergent insights into the factors most strongly associated with survival prediction in this dataset. While traditional, well-established clinical factors like tumor size and patient age were consistently identified as important predictors by the Decision Tree and Random Forest models, the Gradient Boosting model uniquely highlighted the significant influence of lifestyle factors (such as diet and smoking history), specific environmental exposures (like air pollution and factory exposure), particular symptom presentations, certain histology types, and the types of treatment received. These differences in the ranking and set of important features across the models underscore the nature of factors influencing lung cancer survival, where various factors may interact in non-linear ways.

The findings of this study have significant implications for the advancement of precision prevention strategies in lung cancer. The identification of influential factors through machine learning models, even when using readily available clinical and lifestyle data, can serve as powerful indicators of potential underlying genetic vulnerabilities or complex gene-environment interactions that influence survival outcomes. For instance, if specific environmental exposures (such as particular types of air pollution or the use of biomass fuel for cooking/heating) are identified as highly predictive of poorer survival in certain patient subgroups, this could strongly suggest the presence of underlying gene variants that modify an individual's susceptibility or biological response to these specific environmental challenges. Similarly, the observed significance of certain histology types, specific symptom profiles at presentation, or differential responses to various treatment modalities may be linked to the presence of specific oncogenic driver mutations or inherited germline genetic variations that influence tumor biology and treatment sensitivity.

By leveraging machine learning techniques to identify individuals who are predicted to be at a higher risk of poorer survival based on a combination of these influential factors, even in clinical settings where explicit genetic sequencing data may not be readily available, advancement towards implementing more personalized and effective interventions can be done. This paradigm shift allows for tailoring strategies in several key areas:

- **Targeted Screening:** Individuals identified as being at a higher

predicted risk, due to a confluence of unfavorable clinical, lifestyle, and environmental factors (which may indirectly reflect underlying genetic risk), could be prioritized for more frequent or utilize more advanced lung cancer screening methods (such as low-dose CT scans). This aims to detect cancer at an earlier, more curable stage.

- **Personalized Lifestyle Modifications:** Based on the identified important lifestyle and environmental risk factors, personalized and actionable advice can be provided to individuals. This could involve tailored recommendations on reducing exposure to identified environmental hazards or adopting specific dietary habits that research suggests may improve prognosis or mitigate risk in individuals with certain predispositions.
- **Optimized Early Intervention and Treatment Planning:** For individuals predicted to have poorer outcomes based on their risk profile, a more aggressive or highly personalized treatment planning approach can be initiated sooner. This might involve considering novel therapies, clinical trial participation, or multidisciplinary team consultations from the outset.
- **Informed Genetic Counseling and Testing:** The specific combination of identified risk factors and their importance in the predictive models can serve as valuable information to guide decisions about which patients might benefit most from formal genetic counseling and targeted genetic testing. Identifying specific inherited genetic vulnerabilities could further refine risk assessment, inform prophylactic measures for at-risk family members, and guide the selection of highly specific targeted therapies if cancer is diagnosed.

Ultimately, this research demonstrates that machine learning models, even when applied to standard clinical, demographic, and lifestyle data commonly collected in healthcare settings, possess the potential to extract valuable prognostic information. These insights can illuminate the interplay of factors influencing lung cancer survival and, importantly, highlight potential areas where underlying genetic vulnerabilities may be playing a crucial role. This understanding can directly inform the development and implementation of more precise and effective prevention and treatment strategies, thereby paving the way for a truly personalized approach to lung cancer care that moves beyond a one-size-fits-all model.

While this initial study utilized a synthetic dataset, which offers advantages for controlled experimentation and model development, the methodology employed is directly applicable to real-world clinical data. Future work should prioritize the rigorous validation of these findings using diverse real-world clinical datasets from different populations. Ideally, these future studies should incorporate explicit genetic data (such as germline sequencing or somatic mutation profiles) to directly assess and quantify the complex interplay between genetic vulnerabilities, environmental factors, clinical characteristics, and patient survival outcomes. Furthermore, continued model refinement, exploration of more advanced machine learning architectures, and the development of methods to enhance model interpretability will be crucial steps. These advancements will further improve predictive accuracy and clinical utility, ultimately accelerating progress towards more effective precision prevention and personalized medicine strategies in the ongoing fight against lung cancer.

REFERENCE

1. USMAN, M. M., & CHOUDHRY, A. CLINICAL PHARMACY LUNG CANCER ITS PREVALANCE.
2. Hossain, M. A., Asa, T. A., Mahmud, M. Z., Azad, A. K. M., Rahman, M. Z., Moni, M. A., & Moustafa, A. (2025). Genetic Links Between Common Lung Diseases and Lung Cancer Progression: Bioinformatics and Machine Learning Insights. *Emerging Science Journal*, *9*(2), 916-937.
3. Wu, F. Z. (2025). Lung Cancer.
4. Saini, C., Vats, P., Baweja, B., Nirmal, S., & Nema, R. (2025). hsa-let-7b-5p/TMPO-AS1-mediated ceRNA networks are linked to poor prognosis for lung cancer patients with FOXM1/MAD2L1 axis. *Discover Oncology*, *16*(1), 953.
5. Najafiyan, B., Hosseini, Z. B., Esmaelian, S., Firuzpour, F., Anaraki, S. R., Kalantari, L., ... & Nabi-Afjadi, M. (2024). Unveiling the potential effects of resveratrol in lung cancer treatment: Mechanisms and nanoparticle-based drug delivery strategies. *Biomedicine & Pharmacotherapy*, *172*, 116207.
6. Mincuzzi, A., Carone, S., Galluzzo, C., Tanzarella, M., Lagravinese, G. M., Bruni, A., ... & Giannico, O. V. (2024). Gender differences, environmental pressures, tumor characteristics, and death rate in a lung cancer cohort: a seven-years Bayesian survival analysis using cancer registry data from a contaminated area in Italy. *Frontiers in Public Health*, *11*, 1278416.
7. Ochman, B., Kiczmer, P., Ziora, P., Rydel, M., Borowiecki, M., Czyżewski, D., & Drozdzowska, B. (2023). Incidence of Concomitant Neoplastic Diseases, Tumor Characteristics, and the Survival of Patients with Lung Adenocarcinoma or Squamous Cell Lung Carcinoma in Tobacco Smokers and Non-Smokers—10-Year Retrospective Single-Centre Cohort Study. *Cancers*, *15*(6), 1896.
8. Maulini, R., Basyar, M., Herman, D., & Sabri, Y. S. (2025). Beyond the Obstruction: A Case of Lung Cancer with Coincidental COPD Diagnosis. *Bioscientia Medicina: Journal of Biomedicine and Translational Research*, *9*(4), 6844-6856.
9. Yu, F., Xiao, R., Li, X., Hu, Z., Cai, L., & He, F. (2021). Combined effects of lung disease history, environmental exposures, and family history of lung cancer to susceptibility of lung cancer in Chinese non-smokers. *Respiratory research*, *22*(1), 210.
10. Ajayi, R. O., & Ogunjobi, T. T. (2025). Environmental exposures and cancer risk: a comprehensive review. *Medinformatics*, *2*(2), 80-92.
11. Dan, A., Burtavel, L. M., Coman, M. C., Focsa, I. O., Duta-Ion, S., Juganaru, I. R., ... & Radoi, V. E. (2024). Genetic Blueprints in Lung Cancer: Foundations for Targeted Therapies. *Cancers*, *16*(23), 4048.
12. Wang, X., Sharpnack, J., & Lee, T. C. (2025). Improving lung cancer diagnosis and survival prediction with deep learning and CT imaging. *PLoS One*, *20*(6), e0323174.
13. Salmanpour, M. R., Gorji, A., Mousavi, A., Fathi Jouzdani, A., Sanati, N., Maghsudi, M., ... & Rahmim, A. (2025). Enhanced Lung Cancer Survival Prediction Using Semi-Supervised Pseudo-Labeling and Learning from Diverse PET/CT Datasets. *Cancers*, *17*(2), 285.
14. Mahootiha, M., Qadir, H. A., Aghayan, D., Fretland, Å. A., von

Gohren Edwin, B., & Balasingham, I. (2024). Deep learning-assisted survival prognosis in renal cancer: A CT scan-based personalized approach. *Heliyon*, *10*(2).
15. La'ah, A. S., & Chiou, S. H. (2024). Cutting-edge therapies for lung cancer. *Cells*, *13*(5), 436.
16. Zafar, A., Khatoon, S., Khan, M. J., Abu, J., & Naeem, A. (2025). Advancements and limitations in traditional anti-cancer therapies: a comprehensive review of surgery, chemotherapy, radiation therapy, and hormonal therapy. *Discover oncology*, *16*(1), 607.
17. Hua, P., Olofson, A., Farhadi, F., Hondelink, L., Tsongalis, G., Dragnev, K., ... & Hassanpour, S. (2025). Predicting targeted therapy resistance in non-small cell lung cancer using multimodal machine learning. *arXiv preprint arXiv:2503.24165*.
18. Hua PhD, P. (2025). MULTIMODAL MACHINE LEARNING TO ENHANCE PATIENT PROGNOSIS & MANAGEMENT IN CANCER CARE.
19. Adams, S. J., Stone, E., Baldwin, D. R., Vliegenthart, R., Lee, P., & Fintelmann, F. J. (2023). Lung cancer screening. *The Lancet*, *401*(10374), 390-408.
20. Baccarelli, A., Dolinoy, D. C., & Walker, C. L. (2023). A precision environmental health approach to prevention of human disease. *Nature communications*, *14*(1), 2449.
21. Roberts, M. C., Holt, K. E., Del Fiol, G., Baccarelli, A. A., & Allen, C. G. (2024). Precision public health in the era of genomics and big data. *Nature medicine*, *30*(7), 1865-1873.
22. Zaman, M. (2025). Review and importance of precision medicine in cancer treatment and individual health. *Eurasian Journal of Chemical, Medicinal and Petroleum Research*, *4*(2), 237-251.
23. Mahmood, A. A. R., Jha, A. M., & Manivannan, K. (2024). Precision Medicine: Personalizing The Fight Against Cancer. *International Journal of Trends in OncoScience*, 10-18.
24. Schuler, M., Bölükbas, S., Darwiche, K., Theegarten, D., Herrmann, K., & Stuschke, M. (2023). Personalized treatment for patients with lung cancer. *Deutsches Ärzteblatt International*, *120*(17), 300.
25. Alowais, S. A., Alghamdi, S. S., Alsuhebany, N., Alqahtani, T., Alshaya, A. I., Almohareb, S. N., ... & Albekairy, A. M. (2023). Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC medical education*, *23*(1), 689.
26. Shukuya, T., Takahashi, K., Shintani, Y., Miura, K., Sekine, I., Takayama, K., ... & Date, H. (2023). Epidemiology, risk factors and impact of cachexia on patient outcome: results from the Japanese Lung Cancer Registry Study. *Journal of cachexia, sarcopenia and muscle*, *14*(3), 1274-1285.
27. Dacic, S., Travis, W., Redman, M., Saqi, A., Cooper, W. A., Borczuk, A., ... & IASLC Pathology Committee. (2023). International association for the study of lung cancer study of reproducibility in assessment of pathologic response in resected lung cancers after neoadjuvant therapy. *Journal of Thoracic Oncology*, *18*(10), 1290-1302.
28. Erdoğu, V., Çitak, N., Sezen, C. B., Cansever, L., Aker, C., Onay, S., ... & Metin, M. (2021). Correlation between outcomes and tumor size in> 7 cm T4 non-small cell lung cancer patients: A tumor size-based comparison

study. *Asian Cardiovascular and Thoracic Annals*, *29*(8), 784-791.
29. Fan, Y., Jiang, Y., Gong, L., Wang, Y., Su, Z., Li, X., ... & Qiao, Y. (2023). Epidemiological and demographic drivers of lung cancer mortality from 1990 to 2019: results from the global burden of disease study 2019. *Frontiers in public health*, *11*, 1054200.
30. Paappanen, V., Järvenpää, H., Jukkola, A., Päkkilä, P., Sahlström, S., Klaavuniemi, T., ... & Tiainen, S. (2025). Impact of Treatment Decisions on Survival Outcomes in Elderly Patients with Non-Small Cell Lung Cancer: A Retrospective Real-World Study. *Clinical Oncology*, 103930.
31. Islam, M. R., Siddiqua, S. M., Rabbani, G., Al Ayub, S. B., Islam, R., Saha, B., ... & Karim, M. N. (2024). Exploring sex difference in the risk factors and prognosis of inoperable lung cancer. *Cancer Treatment and Research Communications*, *41*, 100848.
32. Emara, H. M. (2026). Targeting the CD47/Calreticulin axis in Triple Negative Breast Cancer using HA grafted Chitosan nanoparticles loaded with Doxorubicin and Polygodial.
33. Wang, X., Romero-Gutierrez, C. W., Kothari, J., Shafer, A., Li, Y., & Christiani, D. C. (2023). Prediagnosis smoking cessation and overall survival among patients with non–small cell lung cancer. *JAMA network open*, *6*(5), e2311966-e2311966.
34. Whitrock, J. N., Carter, M. M., Pratt, C. G., Brokamp, C., Harvey, K., Pan, J., ... & Van Haren, R. M. (2024). The Role of Environmental Exposures on Survival After Non-Small Cell Lung Cancer Resection. *Annals of Thoracic Surgery Short Reports*, *2*(4), 618-623.
35. Mousavi, S. F., Masoudi, S., Rezaei, N., Pourghazi, F., Sharafkhah, M., Eslami, M., ... & Malekzadeh, R. (2025). Survival assessment and pre-diagnostic risk factors for lung cancer incidence: Insights from the Golestan Cohort Study. *PLoS One*, *20*(4), e0320931.
36. Zhou, J., Zheng, Q., Huang, Y., Lyu, M., Wang, T., Wu, D., & Liao, H. (2024). Effect of family history of cancer on postoperative survival in patients with non-small cell lung cancer. *Translational Lung Cancer Research*, *13*(8), 1851.
37. Castro-Espin, C., & Agudo, A. (2022). The role of diet in prognosis among cancer survivors: a systematic review and meta-analysis of dietary patterns and diet interventions. *Nutrients*, *14*(2), 348.
38. Chhatre, S., Vachani, A., Allison, R. R., & Jayadevappa, R. (2021). Survival outcomes with photodynamic therapy, chemotherapy and radiation in patients with stage III or stage IV non-small cell lung cancer. *Cancers*, *13*(4), 803.
39. Garg, A., Iyer, H., Jindal, V., Vashistha, V., Ali, A., Jain, D., ... & Mohan, A. (2022). Prognostic factors for treatment response and survival outcomes after first-line management of Stage 4 non-small cell lung cancer: A real-world Indian perspective. *Lung India*, *39*(2), 102-109.
40. Daniels, J., Kyei, K. A., & Israel, N. (2025). Management and survival outcomes of patients with lung cancer at a leading radiotherapy centre in Sub-Saharan Africa: a cross-sectional study. *ecancermedicalscience*, *19*, 1924.
41. Ballen, D. F., Carvajal-Fierro, C. A., Beltrán, R., Alarcón, M. L., Vallejo-Yepes, C., & Brugés-Maya, R. (2023). Survival outcomes of metastatic non-small cell lung cancer patients with limited access to immunotherapy and

targeted therapy in a cancer center of a low-and middle-income country. *Cancer Control*, *30*, 10732748231189785.
42. Yuan, Q., Cai, T., Hong, C., Du, M., Johnson, B. E., Lanuti, M., ... & Christiani, D. C. (2021). Performance of a machine learning algorithm using electronic health record data to identify and estimate survival in a longitudinal cohort of patients with lung cancer. *JAMA Network Open*, *4*(7), e2114723-e2114723.
43. Torkay, G., Fadlallah, N., Karagöz, A., Canlı, M., Saydam, E., Mete, A., ... & Yeşil, Y. (2024). Artificial intelligence in cancer: a SWOT analysis. *Journal of AI*, *8*(1), 107-137.
44. Muraina, I. (2022, February). Ideal dataset splitting ratios in machine learning algorithms: general concerns for data scientists and data analysts. In *7th international Mardin Artuklu scientific research conference* (pp. 496-504).
45. Hammadi, K. J. (2024). A New Approach for Binary Classification Using Logistic Regression, Decision Trees, and Neural Networks: Evaluation and Comparative Analysis.
46. Ghosh, S. (2024). Comparing regular random forest model with weighted random forest model for classification problem. *International Journal of Statistics and Applications*, *1*, 7-12.
47. Darshith, T. N., & Harshita, P. Comparative Analysis of K-Means Algorithm and Gradient Boosting Algorithm for E-commerce Platform.
48. Dritsas, E., & Trigka, M. (2022). Lung cancer risk prediction with machine learning models. *Big Data and Cognitive Computing*, *6*(4), 139.
49. Janssens, R., Arnou, R., Schoefs, E., Petrocchi, S., Cincidda, C., Ongaro, G., ... & Huys, I. (2021). Key determinants of health-related quality of life among advanced lung cancer patients: a qualitative study in Belgium and Italy. *Frontiers in pharmacology*, *12*, 710518.

MISCELLANEOUS

Link to notebook:
https://colab.research.google.com/drive/1SLZjklG4_ex1lrm6Y4RSrFXYR4bUmaDs?usp=sharing