

Survival Analysis and Prediction of Lung Cancer in Patients based on Clinical and Image Features using Machine Learning

by

Kiran Chhetri

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science (MSc) in Computational Science

The Office of Graduate Studies
Laurentian University
Sudbury, Ontario, Canada

© Kiran Chhetri, 2023

THESIS DEFENCE COMMITTEE/COMITÉ DE SOUTENANCE DE THÈSE
Laurentian Université/Université Laurentienne
Office of Graduate Studies/Bureau des études supérieures

Title of Thesis
Titre de la thèse Survival Analysis and Prediction of Lung Cancer in Patients based on Clinical and Image Features using Machine Learning

Name of Candidate
Nom du candidat Chhetri, Kiran

Degree

Diplôme Master of Science

Department/Program Computational Sciences Date of Defence
Département/Programme Date de la soutenance January 15, 2023

APPROVED/APPROUVÉ

Thesis Examiners/Examineurs de thèse:

Dr. Kalpdrum Passi
(Supervisor/Directeur(trice) de thèse)

Dr. Ratvinder Grewal
(Committee member/Membre du comité)

Dr. Abdel Omri
(Committee member/Membre du comité)

Dr. Mamatha Alugubelly
(External Examiner/Examineur externe)

Approved for the Office of Graduate Studies
Approuvé pour le Bureau des études supérieures
Tammy Eger, PhD
Vice-President Research (Office of Graduate Studies)
Vice-rectrice à la recherche (Bureau des études supérieures)
Laurentian University / Université Laurentienne

ACCESSIBILITY CLAUSE AND PERMISSION TO USE

I, **Kiran Chhetri**, hereby grant to Laurentian University and/or its agents the non-exclusive license to archive and make accessible my thesis, dissertation, or project report in whole or in part in all forms of media, now or for the duration of my copyright ownership. I retain all other ownership rights to the copyright of the thesis, dissertation or project report. I also reserve the right to use in future works (such as articles or books) all or part of this thesis, dissertation, or project report. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that this copy is being made available in this form by the authority of the copyright owner solely for the purpose of private study and research and may not be copied or reproduced except as permitted by the copyright laws without written authority from the copyright owner.

ABSTRACT

Lung cancer develops in lung tissues, most commonly in the cells that line the airways. It is the leading cause of death from cancer in both men and women. To estimate the prevalence of lung cancer in the coming years, it is necessary to diagnose it in the early stages. This thesis work proposes to perform a reliable diagnosis of patients with lung cancer. The goal of this research is to analyze the important variables impacting lung cancer based on p-value using image features as well as clinical data and is focused on quality analysis. Further, to enable early diagnosis of cancer with high efficiency, this work proposes to classify the patient's images into cancer using a Convolutional Neural Network (CNN) to enable its early diagnosis. The thesis discusses the dataset, data pre-processing steps, survival rate risk analysis, classification, and performance evaluation of the process. This study used two kinds of data, clinical and image data. The Genomic Data Commons (GDC) Data Portal and The Cancer Imaging Archive (TCIA) were used as the data source. The Random Forest regression estimation method was used to fill in the missing values. It first imputes all missing data with the mean/mode, then fits a random forest on the observed part and predicts the missing part for each variable with missing values. Three models are used to test the significance of variables on cancer survival rates: Kaplan Meier (KM), Cox Proportional Hazards (CPH), and Accelerated Failure Time (AFT). The analysis took into account three types of data: clinical only, image only, and combined clinical and image data. All three models have been effectively applied and the outcome revealed the most robust data and the crucial variable to be focused upon for further experimentation. For classification, a Convolutional Neural Network (CNN), with low computational cost and time overhead is used. The output of statistical models demonstrates the robustness of image data among all types, as it has the fewest chances of producing false results. Image data, which is common in clinical data collection is less prone to human error. As a result of the data's robustness, only image features data was preferred over clinical data and combined in the

next step to perform the classification of images for cancer prediction. Based on the accuracy, the CNN results were compared to the two other ensemble approaches, Random forest (RF) and XgBoost. CNN achieved an accuracy of 99% in image classification, which was higher than the accuracy rates of Random forest (RF) and XgBoost, which were 95.83% and 95.83%, respectively. As a result, the CNN model can be applied to new Computerized Tomography (CT) scan images for lung cancer diagnosis to conduct additional research and to assist clinicians.

Keywords: Lung Cancer Survival Analysis, Random Forest, XgBoost, Convolutional Neural Network, Genomic Data, The Cancer Imaging Archive, Kaplan Meier, Cox-Proportional hazards model, Accelerated Failure Time Model

ACKNOWLEDGEMENTS

First and foremost, I want to thank God for getting me this far and for blessing me with the right people to assist me at various stages of my studies.

It gives me great pleasure to express my heartfelt gratitude to my thesis advisor, Professor Dr. Kalpdrum Passi, for his encouragement, valuable suggestions, discussion, and guidance throughout my graduate studies. His advice was invaluable throughout the research and writing of this thesis. I couldn't have asked for a better advisor and mentor for my master's programme. This thesis would not have been possible without his guidance and motivation.

I am also grateful to all of my Sudbury friends as well as my Indian friends for their encouragement and assistance in changing my career path. I could not have done it without their assistance.

I'd like to express my heartfelt gratitude and thanks to my loving and caring parents, as well as my dear brothers, for their unwavering support and encouragement throughout my years of study, research, and financial assistance. This achievement would not have been possible without their assistance. Thank you very much.

DEDICATIONS

*This thesis is dedicated to my parents,
For their endless love, support & encouragement.*

TABLE OF CONTENTS

Abstract	iii
Acknowledgment	v
Dedications	vi
List of Tables	ix
List of Figures	x
Abbreviations	xi
Chapter 1	1
Introduction	1
1.1 Lung Cancer and Types.....	5
1.2 Stages of Lung Cancer.....	6
1.3 Approaches for Lung Cancer Survival Analysis.....	8
1.3.1 Statistical Methods.....	9
1.3.2 Classification Methods.....	11
1.4 Need of the Study.....	12
1.5 Problem Statement.....	13
1.6 Proposed Objectives.....	13
1.7 Contributions.....	13-14
1.8 Organization of Thesis.....	14
Chapter 2	15
Literature Survey	15
2.1 Related Work.....	15
2.1.1 Based on Prediction/Classification.....	15
2.1.2 Based on Survival Analysis	20
2.1.3 Combined.....	21
2.2 Research Gaps.....	23
Chapter 3	25
Research Methodology	25
3.1 Dataset Used.....	27
3.2 Data Pre-processing and Cleaning.....	30
3.2.1 Missing Values Treatment.....	31
3.2.2 Association Analysis.....	36
3.3 Survival Analysis Models.....	36
3.3.1 Kaplan-Meier Model Estimate.....	38

3.3.2 Cox proportional hazard model (CPH).....	38
3.3.3 Accelerated failure time model (AFT).....	40
3.4 Classification.....	41
3.4.1 Multilayer CNN.....	42
3.4.2 Random Forest (RF).....	48
3.4.3 XgBoost Model.....	47
3.5 Performance Evaluation and Output.....	49
Chapter 4	52
Results and Discussion	52
4.1 Missing Values Treatment Results.....	52
4.2 Results for Association Analysis.....	54
4.3 Statistical Model Results.....	57
4.3.1 Kaplan-Meier (KM) Model.....	57
4.3.2 Cox Proportional Hazard (CPH) Model.....	70
4.3.3 Accelerated Failure Time (AFT) model.....	76
4.4 Comparison of Models.....	78
4.4.1 CPH Model Comparison with Different Data.....	78
4.5 Classification Results.....	82
4.5.1 Convolutional Neural Network (CNN) Results.....	82
4.5.2 Random Forest (RF) Results.....	83
4.5.3 XgBoost Results.....	84
4.6 Comparison with Previous Studies.....	85
Chapter 5	88
Conclusion and Future Work	88
5.1 Conclusion.....	88
5.2 Future Work and Recommendations.....	89
References	91
Appendix A	98

LIST OF TABLES

Table 1.1:	Estimated new cases and age-standardized incidence rates (excluding Quebec) for cancers by sex in Canada.....	3
Table 1.2:	Estimated deaths and age-standardized mortality rates for cancers by sex in Canada in 2022.....	4
Table 2.1:	Summary of Literature Review	18-20
Table 3.1:	A representation of dataset used.....	27-30
Table 3.2:	Variable with 100% missing values.....	31-32
Table 3.3:	Dataset Statistics.....	44
Table 3.4:	CNN Architecture used in this work.....	45
Table 3.5:	Confusion Matrix.....	50
Table 4.1:	Chi-squared results for association analysis.....	56
Table 4.2:	KM model results.....	66-67
Table 4.3:	CPH model with clinical dataset.....	71
Table 4.4:	CPH model with image dataset.....	73
Table 4.5:	CPH model with combined clinical and image features.....	75
Table 4.6:	Accuracy comparison of the proposed model with state-of-the-art methods..	85-86

LIST OF FIGURES

Figure 1.1	Death proportion in Canada due to cancer and other factors, in 2019.....	2
Figure 1.2	Lung Cancer.....	5
Figure 1.3	AI and its sub-parts.....	11
Figure 3.1	Proposed Methodology.....	26
Figure 3.2	Common Missing Values Plot.....	32
Figure 3.3	Common Missing Values Plot.....	33
Figure 3.4	An example of Data Imputation using Random Forest Model.....	35
Figure 3.5	MissForest Algorithm Pseudocode.....	35
Figure 3.6	Distribution of days_to_death with respect to levels of vital_status.....	36
Figure 3.7	A sample of used image dataset.....	44
Figure 3.8	The basic architecture of CNN.....	44
Figure 4.1	Comparing Distribution of Datasets.....	53
Figure 4.2	Boxplot for outliers.....	54
Figure 4.3	Corrplot to represent correlation.....	55
Figure 4.4	Survival curve of the baseline KM model.....	58
Figure 4.5	KM model with different variables.....	60-62
Figure 4.6	KM model hazard ratio plot.....	67
Figure 4.7	KM model beta coefficient plots.....	68
Figure 4.8	Representation of significant variables as in above table.....	72
Figure 4.9	Output obtained for days_to_death and survival probability using CPH model with clinical dataset.....	72
Figure 4.10	Representation of significant variables as in above table.....	74
Figure 4.11	Output obtained for days_to_death and survival probability using CPH model with image dataset.....	74
Figure 4.12	Representation of significant variables as in above table.....	75
Figure 4.13	Output obtained for days_to_death and survival probability using CPH model with clinical and image dataset.....	77
Figure 4.14	Output obtained for days_to_death and survival probability using AFT model with clinical data.....	77
Figure 4.15	Output obtained for days_to_death and survival probability using AFT model with image data.....	77
Figure 4.16	Output obtained for days_to_death and survival probability using AFT with clinical and image data.....	78
Figure 4.17	Comparison Output.....	79
Figure 4.18	Comparison Output.....	80
Figure 4.19	Comparison Output.....	80
Figure 4.20	Accuracy and Loss results for CNN.....	82
Figure 4.21	Results for CNN.....	83
Figure 4.22	Results for RF.....	84
Figure 4.23:	Results for XgBoost.....	85
Figure 4.24:	Accuracy comparison of the Proposed and previous studies.....	87

ABBREVIATIONS

ASIR	Age-Standardized Incidence Rate
ASMR	Age-Standardized Mortality Rate
NSCLC	Non-small cell lung cancer
SCLC	Small cell lung cancer
AI	Artificial Intelligence
WD	Weibull Distribution
GD	Gumbel Distribution
ED	Exponential Distribution
LD	Log-logistic Distribution
PCM	Partition and clustering methods
KM	Kaplan-Meier
ML	Machine Learning
DL	Deep learning
SVM	Support Vector Machine
KNN	k-Nearest Neighbor
ANN	Artificial Neural Network
CNN	Convolutional Neural Network
LSTM	Long Short Term Memory Network
RNN	Recurrent Neural Networks
DBN	Deep Belief Networks
EL	Ensemble learning
ARM	Association Rule Mining
DT	Decision Tree
ODDN	Optimal Deep Neural Network

CT	Computed Tomography
ILD	Interstitial Lung Disease
KPS	Karnofsky Performance Status
ECOG-PS	Eastern Cooperative Oncology Group Performance Status Scale
TMA	Tissue Microarray
CPH	Cox Proportional Hazards
AFT	Accelerated Failure Time
LUSC	Lung Squamous Cell Carcinoma
GDC	Genomic Data Commons
TCIA	The Cancer Imaging Archive
HR	Hazard Ratio
GLM	Generalized Linear Model
RF	Random Forest
XgBoost	Extreme Gradient Boosting
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
LR	Logistic Regression
DCD-LR	Dual Coordinate Descent method for Logistic Regression
DS	Dataset
DNN	Deep Neural Network
VGG-Net	Visual Geometry Group Network
AUC	Area Under Curve

ANN	Artificial Neural Network
PET	Positron Emission Tomography
LASSO	Least Absolute Shrinkage and Selection Operator
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
NCCTG	The North Central Cancer Treatment Group
NB	Naïve Bayes

Chapter 1

Introduction

In clinical research, there is a constant need to develop effective survival analysis methods for censored data to assess the relationship between risk factors and events of interest. It is widely used to model cancer prognosis to optimize and improve cancer therapy. Cancer is a disease in which certain cells in the body grow uncontrollably and spread to other regions of the body. Cancer can develop almost anywhere in the human body, which contains trillions of cells. Human cells normally grow and multiply (a process known as cell division) to form new cells as the body requires them. Cells die when they become old or damaged, and new cells replace them. When this orderly process fails, abnormal or damaged cells grow and multiply when they should not. These cells can combine to form tumors, which are tissue lumps. Tumors may or may not be cancerous (benign). Cancer is one of the leading causes of death worldwide [1]. About 70% of cancer deaths occur in low and middle-income countries. Cancer is the leading cause of death in Canada, with approximately 1 in 4 women dying from cancer and 1 in 2 women developing cancer in their lifetime [2].

Cancer primarily affects Canadians aged 50 and older, but it can strike anyone at any age. Cancer incidence rates vary across Canada due to differences in risks (including behaviors and exposures) and early detection practices. Similarly, cancer death rates vary due to differences in incidence. Mortality rates may also vary across the country due to differences in access to and outcomes of cancer control activities (for example, screening, diagnosis, treatment, and follow-up). Cancer is the leading cause of death in Canada, accounting for 28.2% of all fatalities. The most common types of cancer in Canada are lung, breast, colorectal, and prostate cancer (excluding non-melanoma skin cancer). In 2023, researchers predicted 239,100 new cancer cases and 86,700 cancer deaths in Canada. (The estimated number of new cases does not include non-melanoma skin cancer cases) [3].

Figure 1.1 gives a pictorial depiction of the death proportion that occurred due to cancer and other causes in Canada as of 2020.

Proportions of deaths due to cancer and other causes, Canada, 2020

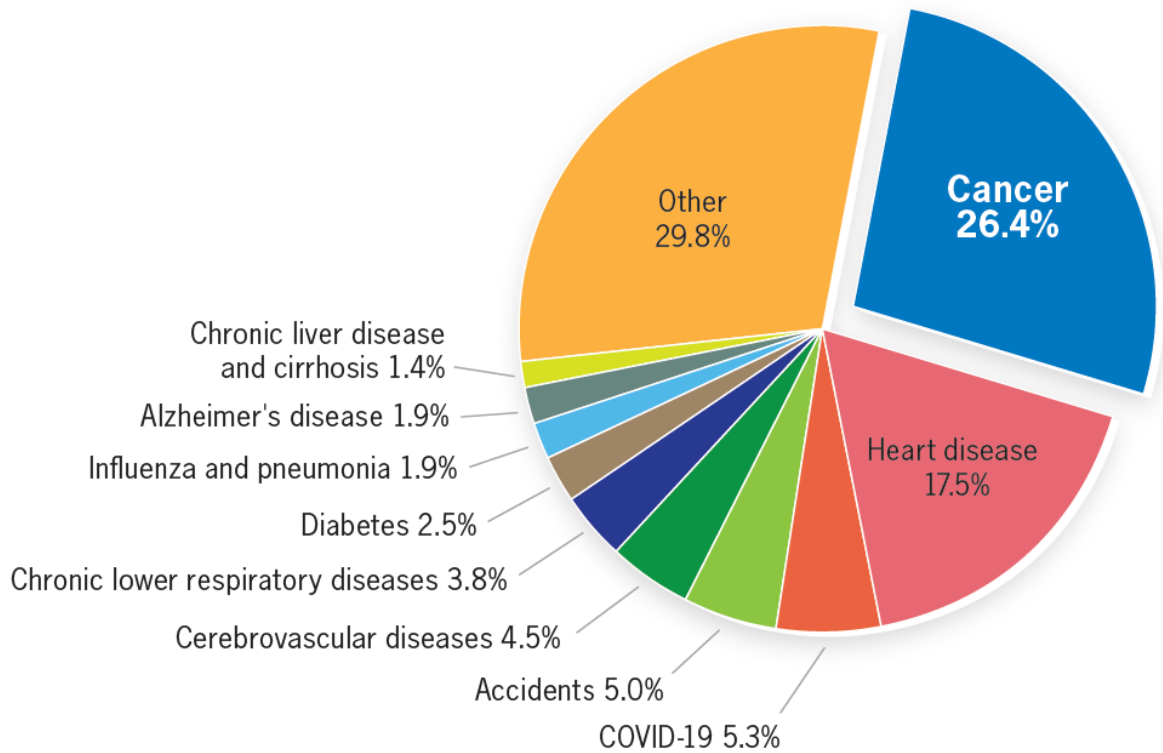


Figure 1.1 Death proportion in Canada due to cancer and other factors, in 2020 [3]

An estimated 239,100 new cancer cases will be expected to be diagnosed in Canada in 2023. The data is presented in Table 1.1. A study by [4] revealed that lung cancer will be the most common cancer diagnosed in Canada, with an estimated 31,000 new cases in 2023. Breast (29,700 cases), prostate (25,900), and colorectal (24,100) cancers are expected to be the next most common cancers. The estimated new cases of these four cancer types represent 46% of all cancers diagnosed in Canada for the year 2023. According to age-standardized incidence rates (ASIRs) (Table 1.1) for Canada, excluding Quebec, males are diagnosed with cancer at a higher rate than females for all non-female-specific cancers except thyroid and breast. Males have a 15% higher ASIR for all cancers combined in 2022 (555.5 M per 100,000 vs 481.2 F per 100,000).

Table 1.1 Estimated new cases and age-standardized incidence rates (excluding Quebec) for cancers by sex in Canada for 2023 [4].

Type of Cancer	New cases (2023 estimates)			Cases per 100,000		
	Total	Males	Females	Both Sexes	Males	Females
All Cancers	239100	124200	114900	513.1	555.3	481.2
Lung and bronchus	31000	15300	15800	58.9	60.1	58.4
Breast	29700	260	29400	68.1	1.2	129.9
Prostate	25900	25900	NA	57.1	120.8	NA
Colorectal	24100	13500	10600	51.1	60.5	42.7
Bladder	13400	10200	3200	25.1	41.9	11.1
Non-Hodgkin lymphoma	10900	6100	4700	24	28.8	19.9
Melanoma	9700	5600	4100	24.3	29.2	20.4
Kidney and renal pelvis	8600	5600	2900	18.6	25.6	12.2
Uterus (body, NOS)	8500	NA	8500	19.6	NA	37.8
Head and neck	7900	5800	2100	16.8	25.7	8.8
Pancreas	7200	4000	3200	14.3	16.8	12
Leukemia	6400	3900	2500	14.3	18.6	10.4
Thyroid	6300	1900	4400	15.9	9.6	22
Liver and intrahepatic bile duct	4700	3200	1450	9.1	13.4	5.1
Stomach	4100	2700	1450	8.5	11.7	5.7
Multiple myeloma	3900	2300	1650	8.2	10.1	6.5
Brain/CNS	3200	1850	1350	7.1	8.6	5.7
Ovary	3100	NA	3100	6.9	NA	13.2
Esophagus	2700	2100	600	5.8	9.6	2.5
Soft tissue (including heart)	1700	950	730	3.9	4.6	3.3
Cervix	1550	NA	1550	4.1	NA	8
Testis	1250	1250	NA	3.3	6.6	NA
Hodgkin lymphoma	1100	640	470	2.7	3.1	2.4
All other cancers	22500	11300	11300	45.5	48.9	43

Further, in 2023, 86,700 Canadians are expected to die from cancer as shown in Table 1.2. Lung cancer is expected to be the leading cause of cancer death, accounting for roughly one-quarter of all cancer deaths (20,600) in Canada. Colorectal cancer (9300), pancreas cancer (5900), breast cancer (5500), and prostate cancer (4900) are projected to be the next leading causes of cancer.

We anticipate that the top five cancer causes will account for 53% of all cancer deaths in Canada by 2023. Lung cancer is the leading cause of cancer death in both men and women (10,800 [23%] and 9,800 [24%] projected deaths, respectively). Colorectal (5200 [11%]), prostate (4,900 [10%]), and

pancreas (3,100 [7%]) are the next leading causes of cancer death in men. Whereas breast cancer (5,400 [13%]), colorectal cancer (4,100 [10%]), and pancreas cancer (2,800 [7%]) are the next leading causes of cancer death in women. Male cancer deaths are expected to be 3% higher than female cancer deaths in 2023. However, the study anticipates that the age-standardized mortality rate (ASMR) for males will be 37% higher than for females (ASMR = 212.3 M per 100,00 vs. 154.6 F per 100,000, respectively) [4]. Therefore, due to the highest prevalence among all, this research work focused on lung cancer to make its prediction using effective techniques.

Table 1.2 Estimated deaths and age-standardized mortality rates for cancers by sex in Canada in 2023 [4]

Type of Cancer	Deaths (2023 estimates)			Deaths per 100,000		
	Total	Males	Females	Both Sexes	Males	Females
All cancers	86700	46500	40200	179.7	212.3	154.6
Lung and bronchus	20600	10800	9800	41.8	48.2	36.9
Colorectal	9300	5200	4100	19.6	24.2	15.7
Pancreas	5900	3100	2800	12.2	13.9	10.7
Breast	5500	55	5400	11.9	0.2	22.1
Prostate	4900	4900	NA	9.8	23	NA
Liver and intrahepatic bile duct ^{†}	3500	2200	1300	7.2	9.9	4.9
Leukemia	3100	1800	1300	6.4	8.3	4.9
Non-Hodgkin lymphoma	3100	1800	1300	6.3	8.2	4.8
Bladder	2600	1850	720	5.2	8.6	2.6
Brain/CNS	2500	1450	1050	5.6	6.8	4.5
Esophagus	2400	1850	550	5	8.3	2.1
Head and neck	2100	1550	580	4.4	7	2.2
Stomach	2000	1250	750	4.2	5.8	2.9
Ovary	1950	NA	1950	4.2	NA	7.9
Kidney and renal pelvis	1900	1250	650	4	5.8	2.4
Multiple myeloma	1700	990	710	3.5	4.5	2.6
Uterus (body, NOS)	1550	NA	1550	3.2	NA	6
Melanoma	1250	820	430	2.7	3.8	1.7
Soft tissue (including heart)	640	350	290	1.4	1.7	1.2
Cervix	400	NA	400	1	NA	1.9
Thyroid	270	120	150	0.6	0.5	0.6
Hodgkin lymphoma	110	70	35	0.2	0.3	0.1
Testis	30	30	NA	0.1	0.2	NA
All other cancers	9300	5000	4300	19.2	23.2	15.9

1.1 Lung Cancer and Types

Lung cancer is a disease that develops as a result of uncontrolled cell division in the lungs. In other words, lung cancer refers to cancers that begin in the lungs, most commonly in the airways (bronchi or bronchioles) or small air sacs (alveoli). The cells divide and replicate themselves as part of their normal function. However, they can experience changes (mutations) that cause them to continue making more of themselves when they shouldn't. Damaged cells that divide uncontrollably form masses of tissue, or tumors that eventually prevent the organs from functioning properly as shown in Figure 1.2 [5].

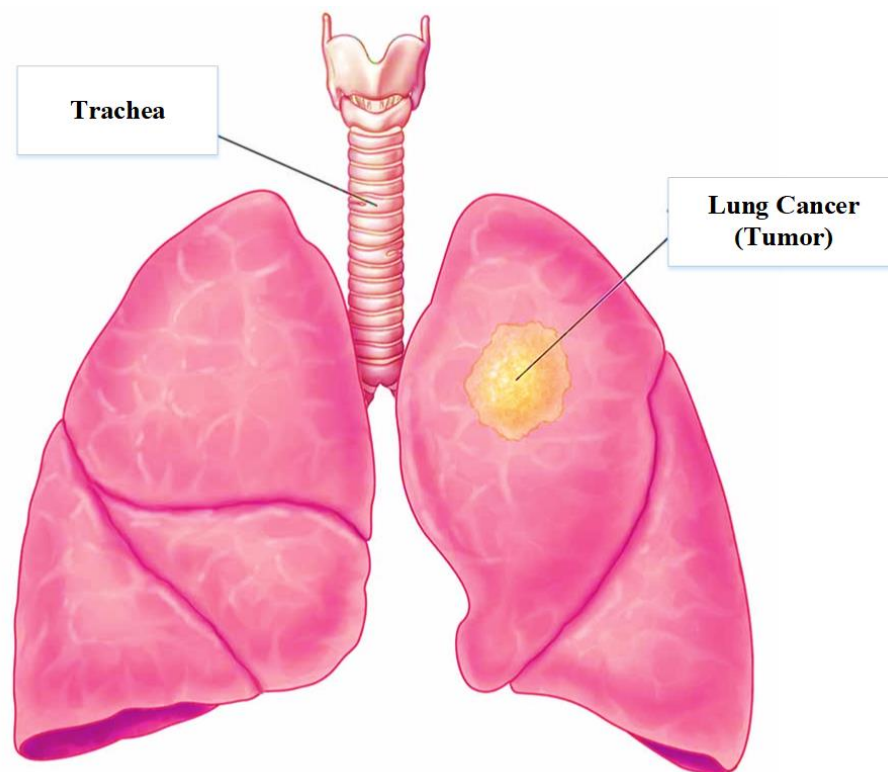


Figure 1.2 Lung Cancer [5]

Many cancers affect the lungs, but the term "lung cancer" usually refers to two types: non-small cell lung cancer and small cell lung cancer [5].

- **Non-small cell lung cancer (NSCLC)**

The most common type of lung cancer is non-small cell lung cancer (NSCLC). It is responsible for more than 80% of lung cancer cases. Adenocarcinoma and squamous cell carcinoma are two common types. Two less common types of NSCLC are adenosquamous carcinoma and sarcomatoid carcinoma.

- **Small cell lung cancer (SCLC)**

Small cell lung cancer (SCLC) grows faster and is more difficult to treat than non-small cell lung cancer (NSCLC). It is frequently discovered as a small lung tumor that has already spread to other parts of your body. Small cell carcinoma (also known as oat cell carcinoma) and combined small cell carcinoma are two types of SCLC.

Other cancers that can begin in or around the lungs include lymphomas (cancer of the lymph nodes), sarcomas (cancer of the bones or soft tissue), and pleural mesothelioma (cancer in the lining of the lungs). These are treated differently and are not typically referred to as lung cancer.

1.2 Stages of Lung Cancer

The size of the initial tumor, how far or deep it penetrates the surrounding tissue, and whether it has spread to lymph nodes or other organs are all factors in cancer staging. Each type of cancer has its staging guidelines. Each stage has several sizes and spread combinations that can fall into that category. For example, the primary tumor in Stage III cancer may be smaller than in Stage II cancer, but other factors may place it at a later stage. Lung cancer is generally staged as follows [6]:

- **Stage 0 (in-situ):** Cancer has spread to the upper lining of the lung or bronchus. It has not spread to other parts of the lung or beyond the lung.
- **Stage 1:** Stage 1 is a number staging system that indicates that your cancer is small. It hasn't reached your lymph nodes or any other distant organs. Stage 1 can be divided into 1A and

1B. The cancer is in stage 1A if it is 3cm or smaller and Stage 1B cancer is between 3cm and 4cm in size. The cancer has not spread beyond the lung.

- **Stage II:** The number staging system includes Stage 2. It is divided into two stages: 2A and 2B. A portion of the affected lung could have collapsed. The cancer is between 4cm and 5cm in size, but there are no cancer cells in any lymph nodes at this stage. Stage 2B indicates that cancer has spread to lymph nodes near the affected lung and is up to 5cm in size. Cancer has spread to lymph nodes within the lung, or multiple tumors exist in the same lobe of the lung.
- **Stage III:** Cancer has spread to nearby lymph nodes or structures, or there are multiple tumors in different lobes of the same lung. Non-small cell lung cancer in stage 3 is also known as locally advanced cancer. Stage 3 is divided into three parts: 3A, 3B, and 3C. Stage 3A can refer to a variety of things. The cancer has spread to the lymph nodes in the center of the chest on the same side as the tumor and has grown up to 5cm in size. Stage 3B can also refer to a variety of things. The cancer is less than 5cm in size and has spread to lymph nodes on the opposite side of the chest from the affected lung, the neck, or above the collarbone. Stage 3C can also refer to a variety of things. The cancer is 5cm to 7cm in size or has spread to one or more of the following areas: the nerve close to the lung (phrenic nerve), the heart's covering (parietal pericardium), and lymph nodes: in the center of the chest, opposite the affected lung, or at the top of the lung, same or opposite side, or above the collar bone.
- **Stage IV:** Cancer has spread to the other lung, lung fluid, the fluid surrounding the heart, or distant organs. There are two stages: 4A and 4B. It's also known as advanced lung cancer. Stage 4A can indicate any of the following: cancer in both lungs, cancer in the covering of the lung (the pleura) or the covering of the heart (the pericardium), fluid around the lungs or the heart containing cancer cells, or a single area of cancer that has spread outside the chest to a lymph node or an organ such as the liver or bone. Stage 4B cancer has spread to multiple areas in one or more organs.

- **Limited vs. extensive stage**

This is determined by whether or not the area can be treated using a single radiation field.

Limited-stage SCLC is confined to one lung and can occasionally be found in lymph nodes in the middle of the chest or above the collar bone on the same side. Extensive stage SCLC has spread throughout one lung or to the opposite lung, lymph nodes on the opposite side of the lung, or other parts of the body.

1.3 Approaches for Lung Cancer Survival Analysis

The time to an event of interest is the primary outcome assessed in many cancer studies. The time is known as survival time, and it can refer to the time 'survived' from complete remission to relapse or progression as well as the time from diagnosis to death. Survival analysis is a subfield of statistics that is also used for modeling or structuring data. The information where the results come in time is also analyzed. The specific difficulties associated with survival analysis stem primarily from the fact that only a subset of the study group has witnessed the event, so survival times will be unknown for a subset of the study group. Censoring is a phenomenon that can occur in the following ways: (a) a patient has not (yet) experienced the relevant outcome, such as relapse or death, by the time the study ends; (b) a patient is lost to follow-up during the study period; or (c) a patient has a different event that makes further follow-up impossible. These censored survival times are an underestimate of the true (but unknown) time to event. When viewing an individual's survival process as a time-line, their event (assuming it occurs) occurs after the end of the follow-up period. This is referred to as right censoring. Censorship can also occur when we observe the presence of a state or condition but are unsure where it originated. Consider a study that looks at the time it takes for cancer to recur after it has been surgically removed. If the patients were examined 3 months after surgery to determine recurrence, those who had a recurrence would have a censored survival time because the recurrence occurred less than 3 months after surgery. Event time data can

also be interval censored, which means that people come in and out of view. If we use the previous example and examine patients at 6 months, those who are disease-free at 3 months and lost to follow-up [7].

Censoring is a feature of survival analysis data that not every cancer patient encounters by the end of their cancer survival time. Censoring is a type of missing or incomplete data. Censoring is present in the models in the same way that it is in the non-parametric hazard and the condition of independent censoring. To perform in-depth analysis for survival analysis of patients with lung cancer, statistical methods and Artificial Intelligence (AI) platforms have been greatly utilized.

1.3.1 Statistical Methods

Statistical methods are mathematical formulas, models, and techniques used in raw research data statistical analysis. Statistical methods are used to extract information from research data and provide various ways to assess the robustness of research outputs. Many researchers use three traditional statistical methods for lung cancer survival analysis: parametric, semi-parametric, and non-parametric as under [7]:

- *Parametric methods*

The distribution of the population from which the sample was drawn is assumed in parametric statistics. When the distribution of survival time is known, this method is used. There are several such models available including Weibull distribution (WD), Gumbel distribution (GD), Exponential distribution (ED), Log-logistic distribution (LD), etc. For fitting survival data, parametric methods involving the exponential, Weibull, lognormal, gamma, and extreme value distributions have been widely used. Gumbel demonstrated that the Weibull and type III smallest extreme value distributions are identical. The log- logistic distribution is an important reliability model because it fits well in many applications of reliability data analysis. Another advantage of the log- logistic distribution is

its closed-form expression for survival and failure rate functions, which distinguishes it from the log-normal distribution [8].

- *Semi-Parametric methods*

A semi-parametric model is a statistical model that includes both parametric and nonparametric components. There are many such models available such as Partition and clustering methods (PCM), Cox proportional hazard model, etc. [9]. Partitioning and clustering methods are semi-parametric approaches for discretizing a multidimensional risk surface into cells with similar risks. However, in the case of such hard clustering methods, the grouping is fixed and sensitive to tuning parameters and initialization, and it can ignore the inherent uncertainty associated with partitioning. The Cox proportional hazards model, on the other hand, is not fully parametric. It is a semi-parametric model because, while the regression parameters (the betas) are known, the outcome distribution is unknown. A Cox model does not specify the baseline survival (or hazard) function. It is inappropriate to use the standard Cox PH model when the PH assumption is not met because it may result in significant bias and loss of power when estimating or inferring the effect of a given prognostic factor on mortality [10].

- *Non-Parametric methods*

The nonparametric method is a type of statistic that makes no assumptions about the sample's characteristics. The Kaplan-Meier (KM) estimate is the most commonly used non-parametric technique for modeling the survival function. Despite all of the difficulties associated with subjects or situations, the Kaplan-Meier estimate is the simplest way to compute survival over time. The Kaplan-Meier survival curve is defined as the probability of survival over a given period when time is divided into many small intervals. In other words, the KM analysis calculates the time it takes from a specific date to death, failure, or other significant events. The product-limit estimator is a non-parametric statistic used to estimate the survival function from lifetime data [11].

1.3.2 Classification Methods

Artificial intelligence (AI) refers to a computer system's ability to mimic human cognitive functions such as learning and problem-solving. A computer system that uses AI uses math and logic to simulate the reasoning that humans use to learn from new information and make decisions. Several concepts come under the AI platform including Machine learning (ML), Deep learning (DL), etc. that can be efficiently used to enhance the diagnostic performance of the system as shown in Figure 1.3 [12].

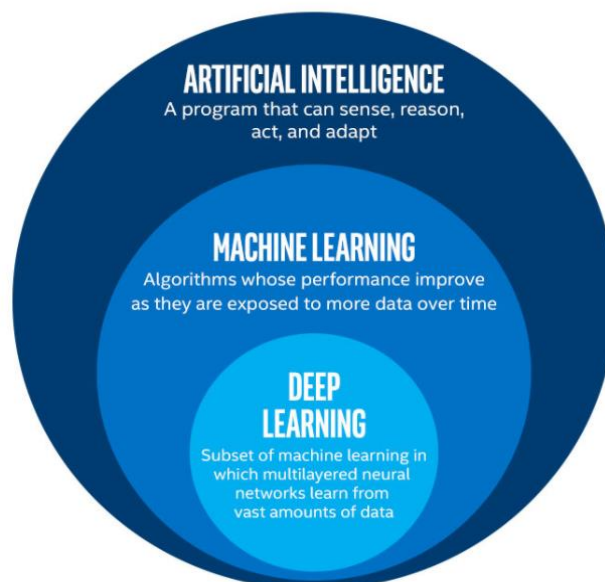


Figure 1.3 AI and its sub-parts [12]

Machine learning is an AI application. It is the process of using mathematical data models to assist a computer in learning without direct instruction. This allows a computer system to learn and improve on its own, based on experience. Machine learning algorithms are classified into four types: supervised, semi-supervised, unsupervised, and reinforcement.

The machine is taught by examples in supervised learning. The operator gives the machine learning algorithm a known dataset with desired inputs and outputs, and the algorithm must figure out how to get those inputs and outputs. The main examples include Support Vector Machine (SVM), k-nearest neighbor (KNN), etc. Semi-supervised learning is similar to supervised learning in that it

employs both labeled and unlabeled data. Labeled data is information that has meaningful tags so that the algorithm can understand it, whereas unlabeled data does not have that information. Machine learning algorithms can learn to label unlabeled data using this combination. Unsupervised learning is a machine learning technique in which models are created without the use of a training dataset. Instead, models discover hidden patterns and insights in the given data, e.g., the K Means Clustering algorithm. Reinforcement learning is concerned with regimented learning processes in which a machine learning algorithm is given a set of actions, parameters, and end values to follow. Following the definition of the rules, the machine learning algorithm attempts to explore various options and possibilities, monitoring and evaluating each result to determine which is optimal, e.g., Artificial Neural Network (ANN) [13][14].

Deep learning (also known as deep structured learning) is a machine learning method that is based on artificial neural networks and representation learning. Deep learning is, in other words, a subset of machine learning in which artificial neural networks, or algorithms inspired by the human brain, learn from large amounts of data. Convolutional Neural Networks (CNNs), Long Short Term Memory Networks (LSTMs), Recurrent Neural Networks (RNNs), Deep Belief Networks (DBNs), and other DL algorithms exist [15].

To increase the efficacy of a system, multiple models are also combined known as Ensemble learning (EL). Ensemble learning is a general meta-machine learning approach that seeks to improve predictive performance by combining predictions from multiple models. Although you can create an infinite number of ensembles for your predictive modeling problem, three methods dominate the field of ensemble learning. Bagging, stacking, and boosting are the three main types of ensemble learning methods [16].

1.4 Need of the Study

Lung cancer is the leading cause of cancer death in men and women alike. The research on lung cancer survival analysis can aid in the development of better treatments, increasing patient survival

and quality of life. This research work can help those diagnosed with lung cancer have a better and longer future, as well as increase the number of survivors living with the disease.

1.5 Problem Statement

To perform survival analysis of patients suffering from lung cancer based on statistical models using clinical, image, and combined data and to predict the same by an effective deep learning technique.

1.6 Proposed Objectives

- To employ a method that can handle crucial missing data values robustly for more efficient diagnosis.
- To evaluate the relevance of the different types of features and data such as image, clinical, and combined (image+clinical) to find the robust one for diagnosis of lung cancer survival.
- To apply an enhanced learning platform such as a deep learning model to classify persons as having cancer and normal for more accurate and early analysis.
- To perform a comparative analysis of several statistical models as well as classification techniques to perform efficient prediction with high accuracy.

1.7 Contributions

- Random Forest based regression estimation method is applied to treat the huge amount of missing data more robustly.
- Kaplan-Meier (KM) model has been utilized to estimate an empirical distribution of survival function.
- The Cox Proportional Hazards (CPH) Model has been employed to perform lung survival analysis.
- The accelerated Failure Time (AFT) Model has been used for modeling the lung survival data.

- An in-depth statistical analysis is performed and the results of the statistical models are analyzed in comparison to find the best one.
- The analysis of lung survival risk is performed by considering the image as well as clinical data to determine their relevancy in such evaluation.
- A deep learning model with less computational overhead has been adopted and employed to perform an effective lung cancer survival analysis using image data and is compared with other models to evaluate its performance in terms of accuracy.

1.8 Organization of Thesis

This thesis is divided into five chapters. Following Chapter 1, the remainder of the thesis is organized as follows. Chapter 2 describes the related work done on the considered area and the gaps identified. Chapter 3 presents the proposed methodology, as well as a detailed explanation of all phases involved. Chapter 4 presents the experimental results for the employed statistical methods and the proposed effective learning model. Also, it provides a discussion of the findings and the analysis made by comparison with other models. Finally, Chapter 5 provides the conclusion as well as future work, followed by references.

Chapter 2

Literature Survey

2.1. Related Work

One of the most important and difficult tasks in Lung Cancer Survival Analysis is model selection. Several research articles were obtained over the years using various models. Many researchers are working in this area to improve the accuracy of the survival risk analysis of lung cancer patients. Before evaluating a model, it is necessary to conduct an in-depth examination and analysis of the state-of-the-art. As a result, this chapter provides an overview of previous work done in the related area of Cancer Survival Analyses using various techniques. The related work is divided into subsections based on the focused task such as survival analysis, classification or prediction or combination of both. The chapter also discusses the problem statement, research gaps discovered, and objectives devised to overcome existing pitfalls in lung cancer analysis.

2.1.1 Related Literature on Prediction/Classification

Palani et al. (2019); continuous monitoring has provided predictive modeling of lung cancer illness. They accomplished this through the use of fuzzy cluster-linked augmentation with categorization. The Otsu thresholding method was used in this study to differentiate the transition area from the lung cancer representation. To achieve incremental classification, the current Association Rule Mining (ARM), conventional decision tree (DT), and CNN are combined with a novel incremental classification technique. Standard images from the database, as well as the most recent data on the patient's health collected from IoT devices attached to the patient, were used to carry out the operations. The study's conclusion indicates that the predictive modeling system has improved in accuracy [18]. Several recent studies considered for the literature survey are given in Table 2.1.

Bhatia et al. (2019); used deep residual learning to develop a method for determining whether or not a CT image contains lung cancer. The researchers created a preprocessing pipeline using the UNet and ResNet models. This pipeline is designed to highlight and extract features from cancerous lung sections. To gather predictions about the likelihood that a CT scan is malignant, an ensemble of XGBoost and random forest classifiers is used. The predictions of each classifier are then pooled, and the final result is used to determine the likelihood of a malignant CT scan. The proposed model has an accuracy that is 84% higher than standard techniques [19].

By reducing the number of characteristics in lung CT scans and comparing it to other classification algorithms, **Lakshmanaprabu et al. (2019)**; developed OODN (Optimal Deep Neural Network). This enabled them to create a more precise method. The use of an automated classification method for lung cancer has reduced the amount of time required for human labeling and eliminated the possibility of errors being made by the person doing the labeling. The researchers discovered that the performance of machine learning algorithms in terms of accuracy and precision in detecting normal and abnormal lung photos has significantly improved. According to the results, the study was successful in classifying lung images with a peer specificity of 94.56%, an accuracy of 96.2%, and a sensitivity of 94.2%. It has been demonstrated that it is possible to improve cancer detection performance in CAT scans [20].

Joon et al. (2019); used an active spline model as their analysis method to segment lung cancer. X-ray images of the lung have been obtained using this technique and X-ray photos. To begin, it is recommended that a median filter be used for noise detection during the preprocessing stage. During the segmentation phase, additional K-means and fuzzy C-means clustering are used to capture feature information. In this study, the final feature retrieval result is obtained after the X-ray image has been segmented. The SVM approach for classification was used to create the recommended model. MATLAB is used to simulate the results of the cancer detection system.

The goal of this study was to detect and classify lung cancer using images that were both normal and malignant [21].

Talukdar et al. (2018); have placed a strong emphasis on the use of image-processing methods for lung cancer diagnosis (2018). Deep learning methodologies are being used to research lung cancer. Lung cancer, the most common type of cancer, claims the lives of an alarmingly large number of people. A computed tomography (CT) scan was used to assess an individual's risk of developing lung cancer. The formation of precancerous tissue is known as "nodules," and their presence is used as a general indicator of cancer. Radiologists with advanced training can detect nodules and often predict their relationship to cancer. These radiologists, however, are capable of producing false positive and false negative results. Because the patient is constantly stressed, a massive amount of data is analyzed, and an appropriate decision for the patient is made promptly. As a result, developing a computer-aided detection system capable of rapidly detecting features based on radiologists' input is most likely the solution [22].

Kurkure et al. (2016); developed a method for the early detection and accurate diagnosis of lung cancer using CT, PET, and X-ray images in 2016 drew a lot of attention and enthusiasm. The use of a genetic algorithm that allows for the early detection of lung cancer nodules by diagnostics allows for the findings to be optimized. To properly and quickly classify the various stages of cancer images, both Naive Bayes and a genetic algorithm were used. This was done to avoid the complexities of the generation process. The categorization has an accuracy rate of up to 80% [23].

Nageswaran et al. (2022); propose an accurate classification and prediction of lung cancer using technology enabled by machine learning and image processing. To begin, photos must be collected. The experimental study used 83 CT scans from 70 different patients as the dataset. During image preprocessing, the geometric mean filter is used. As a result, image quality improves. The images are then segmented using the K-means technique. This segmentation can be used to locate a portion of an image. Then, machine learning classification methods are used. Machine learning techniques

such as ANN, KNN, and RF were used for classification. It has been discovered that the ANN model produces more accurate results [28].

Wang et al. (2018); created an automated tumor region recognition system for lung cancer pathology images using a deep convolutional neural network (CNN). 22 well-defined shape and boundary features were extracted from the identified tumor regions. An independent patient cohort (n=389) was used to develop and validate a tumor region shape-based prognostic model. The predicted high-risk group fared significantly worse than the low-risk group in terms of survival (p value=0.0029). After controlling for age, gender, smoking status, and stage, the predicted risk group serves as an independent prognostic factor (high-risk vs. low-risk, hazard ratio=2.25, 95% CI 1.34-3.77, p value=0.0022). This study adds to our understanding of the relationship between tumor shape and patient prognosis but has certain limitations. First, pathology images capture only a portion of the tumor's characteristics and may not be representative of the entire tumor in terms of size, shape, and other characteristics. Second, the CNN model is sensitive to out-of-focus tissue like red blood cells, macrophages, and stroma cells. Third, image preparation artifacts such as artificially damaged tumor tissues and failure to select images that faithfully represent the tumor can have an impact on image features [31]. Table 2.1 gives the summary of the literature review.

Table 2.1 Summary of Literature Review

Authors (Year)	Study Purpose	Dataset	Technique used	Performance	Conclusion/ Future work
Prediction/Classification					
Nageswaran et al. (2018) [28]	To present an accurate method for classification of lung cancer using ML and image processing	83 CT scans from 70 patients	k-means (segmentation), ANN, KNN, RF (classification)	Accuracy= 99% Sensitivity= 97%	<ul style="list-style-type: none"> • The use of ML can enable the effective diagnosis of lung cancer • KNN is not fit for large dataset • Use of only single DS features • No use of DL platform
She et al. (2020) [35]	To analyze the performance of a deep NN (DeepSurv)	1182 patients with lung cancer	DeepSurv NN	C-statistic: 0.739 vs 0.706 Hazard ratio: 2.99, p<0.01, 95% CI	<ul style="list-style-type: none"> • DL network proved to be promising in prediction of lung cancer • The used models were computationally expensive to train • Lack of external validation
Bebes et al. (2021) [37]	To provide a system that can diagnose different lung cancers efficiently	PET/MR lung images, n=44	SVM, KNN, RF, DL	SVM accuracy= 75.48% (Highest)	<ul style="list-style-type: none"> • The proposed study provides satisfactory results to predict non-small lung cancer • Donot evaluated severity levels

Gu et al. (2019)[38]	To determine the feasibility of ML based radiomics classifier to estimate proliferation of cell in lung cancer analysis	CT scan of 245 patients	Six ML algorithms	RF shown best results: • Sensitivity= 0.762 • Specificity= 0.661	<ul style="list-style-type: none"> • The presented strategy can effectively assess cell proliferation • Only scan data was considered
Survival Analysis					
Zeng et al. (2018) [34]	To propose a non-lab based nomogram with a predictor for analysis of lung cancer clinical features	n= 1106 lung cancer patients	LASSO regression	C-index: 15 days: Training=0.77, Testing=0.78 30 days: Training=0.77, Testing=0.74 90 days: Training=0.75, Testing=0.75	<ul style="list-style-type: none"> • The progression factors identified by LASSO seemed to be useful to make effective nomogram • Retrospective study • Possibility of recall bias • No external validation of research data • Lack of evaluation as no lab indicators were included
Abedi et al. (2014) [24]	To make a comparison of the statistical models to analyze the patients survival rate suffering from lung cancer	n=102 lung cancer patients	Four regression models (Cox, Weibull, Exponential, Gompertz)	Best performance with exponential model AIC, BIC and likelihood: 121.7, 130.3, 13.2	<ul style="list-style-type: none"> • The use of exponential model can make predictions more accurately • Consideration of only single data • No use of ML models for classification
Alomaish et al. (2012) [25]	To analyze whether ILD on CT can impact the survival of patients	146 patients with ILD, 146 lung cancer without ILD	Regression models (Cox and Kaplan Meier)	Death of patients with ILD was 1.522 times without ILD (p=0.029)	<ul style="list-style-type: none"> • Shorter survival rate has been observed in patients with LC and CT with ILD • Retrospective nature of study • Lack of a standardized CT protocol
Mamlook et al. (2020) [27]	To investigate the prevalence of death rate in females and males over time	NCCTG lung cancer dataset	Regression models (Kaplan-Meier, Cox)	For Kaplan= 95% (confidence interval) p-value= 0.0085	<ul style="list-style-type: none"> • Males possess high death risk as compared to females (1.65> females) • No use of ML • Single DS consideration
Deist et al. (2020) [36]	To provide a robust framework which allows data analysis	Lung cancer patient specific databases	A Distributed logistic regression model	-	<ul style="list-style-type: none"> • The proposed infrastructure enables analysis of data from different countries • Lack of use of DL
Prediction and Survival Analysis					
Shayesteh et al. (2020) [32]	To detect the survival time of patients suffering from lung cancer using benefits of radiomics and ML techniques	CT images of 59 patients	Logistic based classification model	DCD-LR and LR achieved best results with an accuracy of 61.02%	<ul style="list-style-type: none"> • The proposed method based on radiomics can be effective for analysis • No use of Deep learning (DL) • Stage of tumor is not considered • Only CT image data is included
Cui et al. (2020) [26]	To propose a survival analysis system using a DL platform	Lung cancer DS (Cancer Genome Atlas)	Deep Neural Network (DNN) based on VGG-Net	• p-value= 0.0045 • c-index= 0.6770	<ul style="list-style-type: none"> • The proposed segmentation free DL technique proved to be useful for lung cancer analysis • More data can be considered in future
Huang et al. (2022) [33]	To analyze scan of lung cancer with ML to predict their progression and overall survival level	n= 965 (PET, CT data)	CNN (Net B4), Random survival forest model (RSF)	For CNN- c-index: • PET+CT, acc.= 79% • AUC= 87% • CT, acc.= 72% • AUC= 88% • PET, acc.= 66% • AUC= 66% For RSF- c-index: • PET+CT+clinical acc.=79% • CT acc.=73% • PET acc.= 59%	<ul style="list-style-type: none"> • The combination of data can yield better results in estimation of survival value of lung cancer patients • No severity level consideration • More number of clinical features can be included • Less generalizability

Wang et al. (2021) [29]	To propose a novel method to perform classification of survival and Cox regression	1137 patients with non-small lung cancer	SurvNet (multi-task based NN)	C-index: Cox model:0.5955 Cox net: 0.5617 SurvNet: 0.6003	<ul style="list-style-type: none"> • The use of SurvNet can evaluate the prediction accurately to diagnose cancer • No use of DL • Evaluated only single DS
Yu et al. (2016) [30]	To present a system that can extract features useful to predict lung cancer	TCGA and Tissue Microarray DS	RF, SVM, VB, Bagging	p<0.036	<ul style="list-style-type: none"> • The proposed strategy can predict the lung cancer prognosis • No use of ML
Wang et al. (2018) [31]	To propose a DL based network and building a Cox proposition hazards model for prediction of survival time	<ul style="list-style-type: none"> • 274 patients CT images without survival time • 129 patients with survival time 	Residual Convolutional auto model	C-index: • DL features= 0.70 • Hand crafted= 0.62	<ul style="list-style-type: none"> • The proposed DL approach and the regression model can effectively predict survival rate of lung cancer • The combination of clinical and non-clinical data is not evaluated • Lack of severity levels consideration

2.1.2 Related Literature on Survival Analysis

Abedi et al. (2019); conducted a study to compare various statistical models of survival and to determine the survival rate and its associated factors among lung cancer patients. The cumulative survival rate, median survival time, and factors associated with lung cancer survival were estimated in this retrospective cohort using Cox, Weibull, exponential, and Gompertz regression models. The log-rank test and Kaplan-Meier tables were also used to examine patient survival in different subgroups. The hazard ratios (95% CIs) for male sex, age, and SCLC were 0.56 (0.33 to 0.93), 1.03 (1.01 to 1.05), and 2.91 (1.71 to 4.95), respectively, according to multivariate analyses. The results demonstrated that the exponential model was the most accurate. This model identified age, gender, and cancer type as factors that predicted survival in lung cancer patients [24].

The study by **Alomaish et al. (2021)**; sought to determine whether the presence of ILD documented on CT affects the survival of lung cancer patients. The study included 146 patients who had interstitial lung disease (ILD) at the time of their initial chest CT. Chest CTs were analyzed to determine the presence of pulmonary fibrosis, which was classified into four types. The presence and type of emphysema, the extent of ILD and emphysema, the location and histologic type of cancer, clinical staging, and treatment were all considered. Different groups' survival probability and hazard of death were assessed using Kaplan-Meier estimates and Cox regression models. P 0.05 was considered significant. There was no significant difference in survival rate between the four different

categories of ILD (log-rank test, $p = 0.195$) or histologic types (log-rank test, $p = 0.4005$). A cox proportional hazard model with ILD, clinical stage, and age was used. The risk of death was 1.522 times higher in patients with ILD than in patients without ILD (95% CI, $p = 0.029$). According to the study, patients with lung cancer and CT evidence of ILD have significantly shorter survival than patients with only lung cancer [25].

Mamlook et al. (2020); conducted a study to perform survival analysis and assess the performance scores of patients with advanced lung cancer based on daily activities. A scoring system was developed using data from 228 patients. Cox's proportional hazards regression and the Kaplan-Meier method were used for survival analysis. The Karnofsky Performance Status (KPS) and the Eastern Cooperative Oncology Group Performance Status Scale (ECOG PS) are used by physicians to assess their ability to perform their usual daily activities. According to the study's findings, patients with scores greater than zero face an increased risk of death. Because there are no significant increased risks of death for patients, the KPS score is not a good predictor of survival for advanced lung cancer patients. However, because the ECOG score was a good predictor of survival, it is refreshing that doctors can use it as a tool to help them make more informed treatment decisions [27].

2.1.3 Related Literature on both Prediction and Survival Analysis

Cui et al. (2020); proposed a survival analysis system that makes use of recently developed deep learning techniques. The proposed system is made up of three major parts. 1) The first component is a deep neural network-based cellular feature learning module with global average pooling. 2) For robust feature selection and survival analysis, the second component is a Cox proportional hazards model with an elastic net penalty. 3) The third component is a biomarker interpretation module, which can aid in the localization of image regions that influence the survival model's decision. Extensive experiments show that the proposed survival model has excellent predictive power in terms of two commonly used metrics for a public (i.e., The Cancer Genome Atlas) lung cancer dataset: log-rank test (p-value) of the Kaplan-Meier estimate and concordance index (c-index). The

method used to visualize the discovered biomarkers, can be used as concrete evidence to back up the survival model's decision [26].

SurvNet, a novel multi-task-based neural network, is proposed in a study by **Wang et al. (2021)**. SurvNet is trained in a multi-task learning framework to learn jointly across three related tasks: input reconstruction, survival classification, and Cox regression. Furthermore, the SurvNet model uses context gating to bridge the gap between survival classification and Cox regression. To assess the performance of the SurvNet model, a new real-world dataset of 1,137 patients with IB-IIA stage non-small cell lung cancer is collected. SurvNet outperforms the traditional Cox model and Cox-Net in terms of concordance index. SurvNet's difference between high-risk and low-risk groups is greater than the difference between high-risk and low-risk groups obtained by the other models. Furthermore, even when the input data is randomly cropped, the SurvNet outperforms the other models, and it achieves better generalization performance on the SEER dataset [29].

Wang et al. (2014); proposed a novel image marker-based integrated framework for non-small cell lung cancer (NSCLC) computer-aided diagnosis and survival analysis. To accurately segment each cell in digital images, a robust seed detection-guided cell segmentation algorithm is proposed. Following that, eight different classification techniques capable of handling high-dimensional data were evaluated and compared for computer-aided diagnosis. The results show that random forest and AdaBoost provide the best NSCLC classification performance. Finally, component-wise likelihood-based boosting is used to fit a Cox proportional hazards model. The survival analysis demonstrates that the statistical model developed from the discovered image markers has a high predictive power [17].

Yu et al. (2016); used The Cancer Genome Atlas (TCGA) to obtain 2,186 hematoxylin and eosin-stained histopathology whole-slide images of lung adenocarcinoma and squamous cell carcinoma patients, and the Stanford Tissue Microarray (TMA) Database to obtain 294 additional images. They extracted 9,879 quantitative image features from the TCGA data set and used regularised machine-

learning methods to select the top features and differentiate shorter-term survivors from longer-term survivors with stage I adenocarcinoma ($P = 0.003$) or squamous cell carcinoma ($P = 14\ 0.023$). The findings suggest that automatically derived image features can predict lung cancer prognosis and thus contribute to precision oncology. Their methods can be applied to histopathology images from other organs [30].

2.2 Research Gaps

Based on the state-of-the-art literature studied and the findings obtained, there are some research gaps identified that need to be focused on and covered to improve the efficacy of the diagnostic system for robust lung cancer survival analysis:

- The literature reveals the use of only single data alone to perform lung cancer survival analysis despite evaluating the reliability of other types of data to find the best one.
- Very little use of effective learning platforms such as deep learning is seen in previous studies. Most of the studies just adopted Machine learning models that may have low-performance rates.
- Non-consideration of robust methods to handle the issue of missing feature values.
- A few numbers of features were considered in the literature and lacked in-depth statistical analysis.

The proposed approach considered multimodal data including clinical, images and combination of both to perform survival analysis which lacks in the past work. Most of the past studies only considered either clinical or image data and not evaluated the advantages of combining both of them. Further, statistical and DL based analysis is not used much by the existing works and needs. Some of the researcher's undertaken only statistical analysis while other focused on only prediction. The proposed approach focuses on both statistical analysis and prediction of patients suffering from lung cancer and makes use of efficient techniques to investigate the best one With the implication of proposed strategy and

techniques, this work has successfully enhanced the accuracy of existing models, thus providing overall best prediction results.

Chapter 3

Research Methodology

In this work, we proposed a novel methodology to perform a risk analysis of the survival rate of patients suffering from lung cancer using statistical methods as well as advanced learning techniques. Figure 3.1 depicts the proposed workflow used and implanted for such analysis. The proposed framework consists of different phases including Dataset used, Data pre-processing and cleaning, Survival rate risk statistical analysis, classification, and performance evaluation. Each phase is explained in detail below.

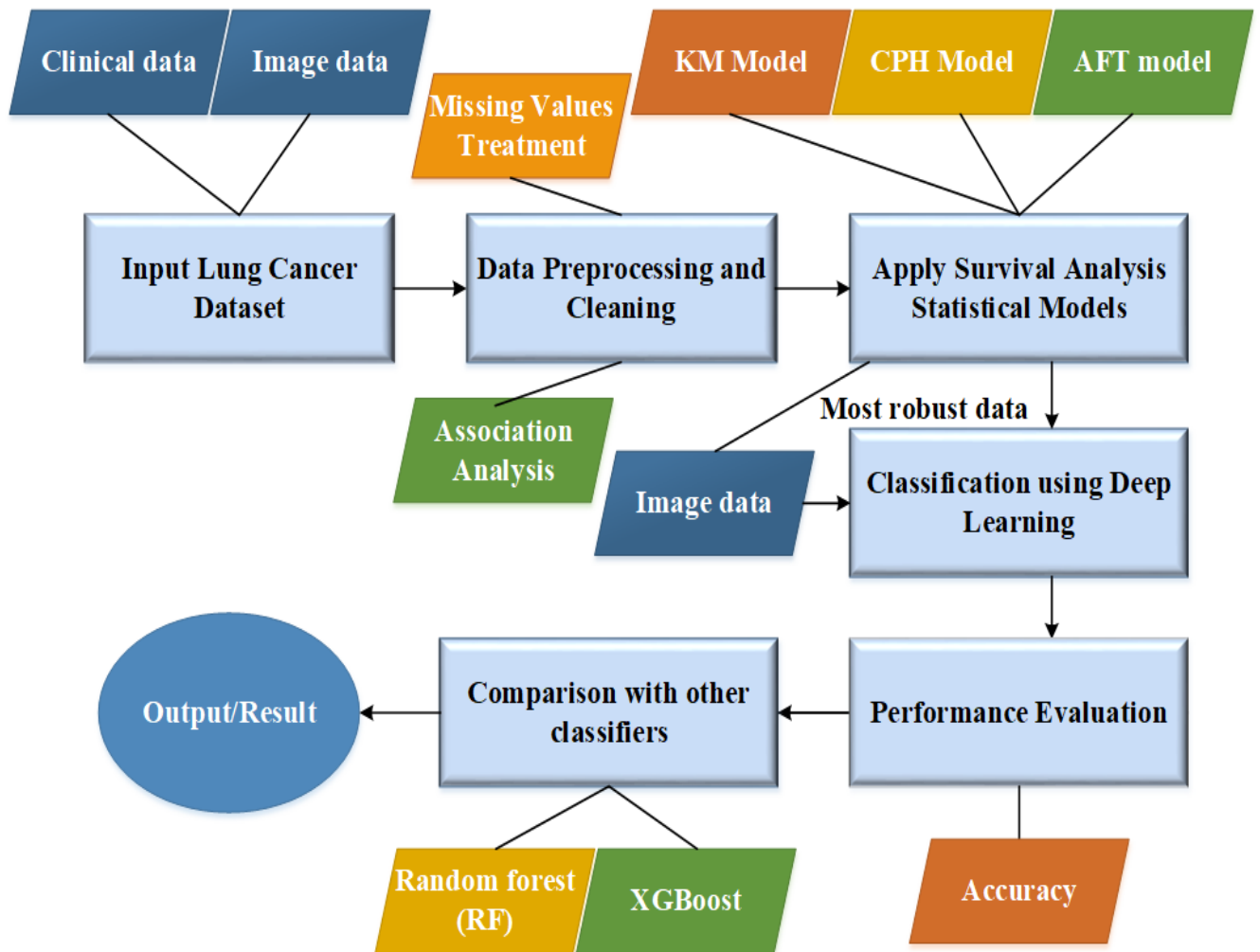


Figure 3.1 Proposed Methodology

3.1 Dataset Used

This work used two types of data i.e. clinical as well as image data collected from [39]. The collection of data from The Cancer Genome Atlas Lung Squamous Cell Carcinoma (TCGA-LUSC) is part of a larger effort to build a research community focused on connecting cancer phenotypes to genotypes by providing clinical images matched to subjects from The Cancer Genome Atlas (TCGA). The Genomic Data Commons (GDC) Data Portal houses clinical, genetic, and pathological data, while The Cancer Imaging Archive houses radiological data (TCIA). The images were obtained in the majority of cases as part of routine care rather than as part of a controlled research study or clinical trial.

There were 171 variables and 504 observations in the data. Manifest variables include days to death and vital status (required for survival analysis); 504 vectors with 64 columns representing image pixel values; and 105 clinical variables. Each variable presents a unique meaning necessary for survival analysis. The image data is first converted into numerical form using a grayscale method. The method scales all values between 0 and 1 and converts them into a single vector form. A screenshot representing the variables' names and image data is presented in Table 3.1.

Table 3.1. A representation of the dataset

```
[1] "bcr_patient_uuid"
[2] "bcr_patient_barcode"
[3] "form_completion_date"
[4] "tissue_prospective_collection_indicator"
[5] "gender"
[6] "diagnosis"
[7] "ethnicity"
[8] "tissue_retrospective_collection_indicator"
[9] "year_of_tobacco_smoking_onset"
[10] "stopped_smoking_year"
[11] "eastern_cancer_oncology_group"
[12] "karnofsky_performance_score"
[13] "race"
[14] "laterality"
[15] "other_dx"
[16] "anatomic_neoplasm_subdivision"
[17] "history_of_neoadjuvant_treatment"
[18] "location_in_lung_parenchyma"
```

[19] "histological_type"
[20] "year_of_initial_pathologic_diagnosis"
[21] "residual_tumor"
[22] "system_version"
[23] "pathologic_T"
[24] "initial_pathologic_diagnosis_method"
[25] "pathologic_N"
[26] "init_pathology_dx_method_other"
[27] "pathologic_M"
[28] "days_to_last_followup"
[29] "pathologic_stage"
[30] "clinical_T"
[31] "clinical_N"
[32] "clinical_M"
[33] "days_to_death"
[34] "clinical_stage"
[35] "pulmonary_function_test_performed"
[36] "pre_bronchodilator_fev1_percent"
[37] "post_bronchodilator_fev1_percent"
[38] "pre_bronchodilator_fev1_fvc_percent"
[39] "post_bronchodilator_fev1_fvc_percent"
[40] "kras_gene_analysis_performed"
[41] "dlco_predictive_percent"
[42] "kras_mutation_found"
[43] "kras_mutation_result"
[44] "egfr_mutation_performed"
[45] "egfr_mutation_result"
[46] "eml4_alk_translocation_performed"
[47] "eml4_alk_translocation_result"
[48] "eml4_alk_translocation_method"
[49] "tobacco_smoking_history"
[50] "vital_status"
[51] "number_pack_years_smoked"
[52] "performance_status_scale_timing"
[53] "radiation_therapy"
[54] "postoperative_rx_tx"
[55] "primary_therapy_outcome_success"
[56] "person_neoplasm_cancer_status"
[57] "new_tumor_event_after_initial_treatment"
[58] "days_to_birth"
[59] "age_at_initial_pathologic_diagnosis"
[60] "age_began_smoking_in_years"
[61] "anatomic_neoplasm_subdivision_other"
[62] "cd4_counts_at_diagnosis"
[63] "cdc_hiv_risk_group"
[64] "days_to_hiv_diagnosis"
[65] "days_to_initial_pathologic_diagnosis"
[66] "days_to_patient_progression_free"
[67] "days_to_sample_procurement"
[68] "days_to_tumor_progression"
[69] "death_cause_text"
[70] "disease_code"
[71] "extranodal_involvement"
[72] "hbv_test"

[73] "hcv_test"
[74] "height"
[75] "history_immunological_disease"
[76] "history_immunological_disease_other"
[77] "history_immunosuppressive_rx"
[78] "history_immunosuppressive_rx_other"
[79] "history_of_other_malignancy"
[80] "history_relevant_infectious_dx"
[81] "history_relevant_infectious_dx_other"
[82] "hiv_rna_load_at_diagnosis"
[83] "hiv_status"
[84] "hvp_test"
[85] "icd_10"
[86] "icd_o_3_histology"
[87] "icd_o_3_site"
[88] "informed_consent_verified"
[89] "kshv_hhv8_test"
[90] "lost_follow_up"
[91] "lymph_node_examined_count"
[92] "nadir_cd4_counts"
[93] "number_of_lymphnodes_positive_by_he"
[94] "number_of_lymphnodes_positive_by_ihc"
[95] "on_haart_therapy_at_cancer_diagnosis"
[96] "on_haart_therapy_prior_to_cancer_diagnosis"
[97] "patient_death_reason"
[98] "patient_id"
[99] "pos_lymph_node_location"
[100] "pos_lymph_node_location_other"
[101] "primary_lymph_node_presentation_assessment"
[102] "prior_aids_conditions"
[103] "project_code"
[104] "stage_other"
[105] "tissue_source_site"
[106] "tumor_tissue_site"
[107] "weight"
[108] "X"
[109] "X.1"
[110] "X.2"
[111] "X.3"
[112] "X.4"
[113] "X.5"
[114] "X.6"
[115] "X.7"
[116] "X.8"
[117] "X.9"
[118] "X.10"
[119] "X.11"
[120] "X.12"
[121] "X.13"
[122] "X.14"
[123] "X.15"
[124] "X.16"
[125] "X.17"
[126] "X.18"

[127]	"X.19"		
[128]	"X.20"		
[129]	"X.21"	[161]	"X.53"
[130]	"X.22"	[162]	"X.54"
[131]	"X.23"	[163]	"X.55"
[132]	"X.24"	[164]	"X.56"
[133]	"X.25"	[165]	"X.57"
[134]	"X.26"	[166]	"X.58"
[135]	"X.27"	[167]	"X.59"
[136]	"X.28"	[168]	"X.60"
[137]	"X.29"	[169]	"X.61"
[138]	"X.30"	[170]	"X.62"
[139]	"X.31"	[171]	"X.63"
[140]	"X.32"		
[141]	"X.33"		
[142]	"X.34"		
[143]	"X.35"		
[144]	"X.36"		
[145]	"X.37"		
[146]	"X.38"		
[147]	"X.39"		
[148]	"X.40"		
[149]	"X.41"		
[150]	"X.42"		
[151]	"X.43"		
[152]	"X.44"		
[153]	"X.45"		
[154]	"X.46"		
[155]	"X.47"		
[156]	"X.48"		
[157]	"X.49"		
[158]	"X.50"		
[159]	"X.51"		
[160]	"X.52"		

In Table 3.1, data from [1]-[107] indicates the clinical demographic data related to each patient ID. For example, [5], [34], [50], and [74] represent the gender (either male or female), clinical stage (early, moderate, or severe), vital status (dead or alive), and height of the subject. Similarly, other clinical data is recorded for each subject to perform detailed analysis. The data from [108]-[171] indicate the image pixel values data used for this experimentation purposes.

3.2 Data Pre-processing and Cleaning

Data preprocessing is a critical task. It's a data mining technique that turns raw data into something more understandable, useful, and efficient. Preprocessing includes data cleaning, which is used to

handle many irrelevant and missing parts of data. Scrubbing is another term for it. This task entails filling in missing values, smoothing or removing noisy data, and resolving inconsistencies. In this work, to perform preprocessing of data, two main operations are performed i.e. missing values treatment and association analysis. Basic exploratory data analysis is performed to better understand the data. Summary statistics, missing value analysis, and the distribution of dependent and independent variables were performed first.

3.2.1 Missing Values Treatment

The analysis of the dataset revealed that it contains 33077 missing cells, accounting for 38.38% of the total number of cells in the dataset. In missing value treatment, group-wise mean/most common values are imputed. Mode and median can also be used to impute missing values as per the type of variable. The variables with a high proportion of missing values need to be eliminated from the data. The following variables as presented in Table 3.2 contained 100% missing values hence these variables need to be focused and removed. To analyze the missing variables more precisely, a plot is drawn and presented in Figure 3.2.

Table 3.2. Variable with 100% missing values

```
[1] "laterality"
[2] "initial_pathologic_diagnosis_method"
[3] "init_pathology_dx_method_other"
[4] "clinical_T"
[5] "clinical_N"
[6] "clinical_M"
[7] "clinical_stage"
[8] "kras_mutation_found"
[9] "kras_mutation_result"
[10] "egfr_mutation_result"
[11] "eml4_alk_translocation_result"
[12] "eml4_alk_translocation_method"
[13] "age_began_smoking_in_years"
[14] "anatomic_neoplasm_subdivision_other"
[15] "cd4_counts_at_diagnosis"
[16] "cdc_hiv_risk_group"
[17] "days_to_hiv_diagnosis"
[18] "days_to_patient_progression_free"
[19] "days_to_sample_procurement"
[20] "days_to_tumor_progression"
[21] "death_cause_text"
```

```

[22] "disease_code"
[23] "extranodal_involvement"
[24] "hbv_test"
[25] "hcv_test"
[26] "height"
[27] "history_immunological_disease"
[28] "history_immunological_disease_other"
[29] "history_immunosuppressive_rx"
[30] "history_immunosuppressive_rx_other"
[31] "history_of_other_malignancy"
[32] "history_relevant_infectious_dx"
[33] "history_relevant_infectious_dx_other"
[34] "hiv_rna_load_at_diagnosis"
[35] "hiv_status"
[36] "hpv_test"
[37] "kshv_hhv8_test"
[38] "lost_follow_up"
[39] "lymph_node_examined_count"
[40] "nadir_cd4_counts"
[41] "number_of_lymphnodes_positive_by_he"
[42] "number_of_lymphnodes_positive_by_ihc"
[43] "on_haart_therapy_at_cancer_diagnosis"
[44] "on_haart_therapy_prior_to_cancer_diagnosis"
[45] "patient_death_reason"
[46] "pos_lymph_node_location"
[47] "pos_lymph_node_location_other"
[48] "primary_lymph_node_presentation_assessment"
[49] "prior_aids_conditions"
[50] "project_code"
[51] "stage_other"
[52] "weight"

```

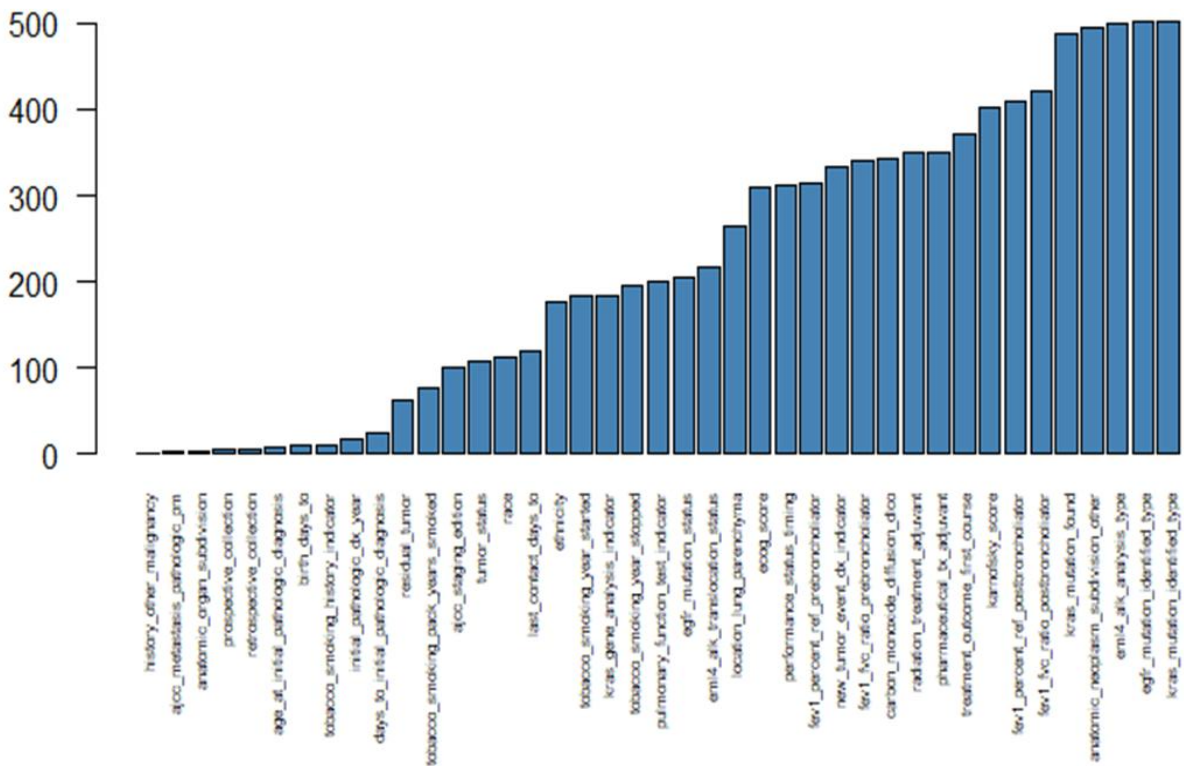


Figure 3.2. Common Missing Values Plot

As it is clear from Table 3.2, the "days to death" variable has a very high proportion of missing values; as it is an important variable for survival analysis, we need to impute it with a more robust method. Simply replacing it with mean, mode, and median would not make it representative in such cases. 294 of the 504 observations have NA values in the "days to death" variable. This variable is necessary for survival analysis. There are two options: remove all NAs and perform the analysis with the remaining observations, replacing the NA values with group means calculated on various other factor variables such as Gender, Ethnicity, Race, duration of smoking, pathological stage, and several pack years smoked, or use an automated robust algorithm to fill the missing values to avoid biases and errors due to human intervention and judgment. To calculate the duration_of_smoking variable, the year_of_tobacco_smoking_onset variable is deducted from the stopped_smoking_year variable. Gender-race and gender-ethnicity have common missing values. After converting in years (days to birth/365), the variable days to birth must be divided into two levels: age less than or equal to 70 years and age greater than 70 years. The break-point of 70 was chosen because the patient's median and mean age are both around 70, as shown in Figure 3.3.

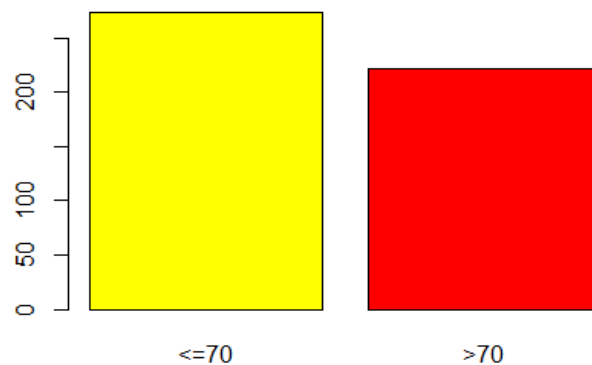


Figure 3.3 Common Missing Values Plot

Other variables that were not more relevant are also removed. To use any machine learning algorithm, the variables must be in numeric, factor, or integer form. There are many variables in the character form in the dataset. The zero variance variables have constant values throughout the observations and do not add value in model building. We removed zero variance and non-zero variance variables. The indexes were removed because they contributed nothing to the model. After

completing these data filtering steps, the dataset is ready for imputing days to death variable values using the random forest regression. The Random Forest regression estimation method [40] was used, with the set of independent variables (with zero or very low missing values) and days to death (observations with no missing values) as the dependent variables. The observations with no missing days_to_death values were used to train the regression model, and the observations with missing days_to_death values were used to predict the days-to-death values and then filled.

MissForest [41] is a missing data imputation algorithm implemented in R in the missForest() package. MissForest first imputes all missing data with the mean/mode, then for each variable with missing values, it fits a random forest on the observed part and predicts the missing part. In other words, MissForest is a machine learning-based data imputation algorithm that is similar to the Random Forest algorithm. The algorithm's creators, Stekhoven and Buhlmann, published a study in 2011 comparing imputation methods on datasets with randomly introduced missing values. MissForest outperformed all other algorithms, including KNN-Impute, in all metrics, in some cases by more than 50%. First, the missing values are filled in using median/mode imputation. The missing values are then labeled as 'Predict,' while the others serve as training rows for a Random Forest model trained to predict, in this case (as shown in Figure 3.4), Age based on Score. The predicted value for that row is then filled in to create a transformed dataset. This looping through missing data points process is repeated several times, with each iteration improving on better and better data. It's like standing on a pile of rocks and constantly adding more to raise yourself: the model uses its current position to raise itself even higher.

Iterations continue until a stopping criterion is met or a predetermined number of iterations have elapsed. In general, datasets are well-imputed after four to five iterations, but this varies depending

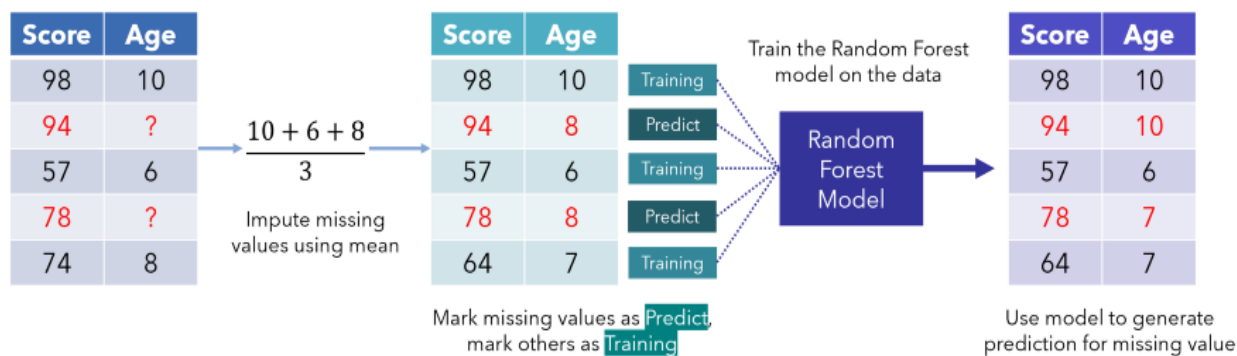


Figure 3.4 An example of Data Imputation using the Random Forest Model

on the size and amount of missing data. The pseudocode for the MissForest Algorithm is presented in Figure 3.5. The sanity of the data was checked after MissForest imputation using the random forest algorithm. We also examined the distribution of the days_to_death variable in data1 and data_impute_dataframes using a histogram. If the distributions of the two datasets are similar, we can conclude that they are from the same population.

Requires:

- X , our $n \times p$ matrix for imputation
- γ , the stopping criterion (it works in iterations, stopping when the difference between iteration i and $i + 1$ of the imputed dataframes begins to *increase* for categorical *and* numeric variables, or after 10 iterations with the default `maxiter` parameter)

Algorithm:

1. Make an **initial guess** for all missing categorical/numeric values (e.g. mean, mode)
2. $k \leftarrow$ vector of column indices in X , sorted in **ascending order of % missing**
3. **while** not γ **do**:
4. $X_{old}^{imp} \leftarrow$ store previous imputed matrix
5. **for** s in k **do**:
6. Fit a random forest predicting the non-missing values of X_s : $y_{obs}^{(s)} \sim x_{obs}^{(s)}$
7. Use this to predict the missing values of X_s : predict $y_{mis}^{(s)}$ using $x_{mis}^{(s)}$
8. $X_{new}^{imp} \leftarrow$ update imputed matrix, using the predicted $y_{mis}^{(s)}$
9. **end for**
10. update γ
11. **end while**
12. **return** the final imputed matrix X^{imp}

Figure 3.5 MissForest Algorithm Pseudocode

There are numerous advantages to using **MissForest**. This method is best suitable for mixed type of data. So, it can be used with both numerical and categorical data and no pre-processing is required.

Random Forest can handle these aspects of data because it does not make feature relationship assumptions like K-Nearest Neighbors does. Because Random Forest is non-parametric, no tuning is required and it is robust to outliers. It can also work with high-dimensional data and is immune to the dimensional curse.

3.2.2 Association Analysis

It is critical to comprehend the relationship between the variables. If two variables are discovered to be highly associated, one of them must be removed from the dataset to avoid the effects of linear combatants during the model-building process. The results obtained are presented in Chapter 4.

3.3 Survival Analysis Models

For survival analysis, we need survival rate and event data as manifest variables. In the dataset for lung cancer diagnosis, the `days_to_death` variable (measured in days) is the survival rate and the `vital_status` variable (binary variable having two levels - Living and Dead) is an event. Figure 3.6 shows the distribution of `days_to_death` concerning levels of `vital_status`.

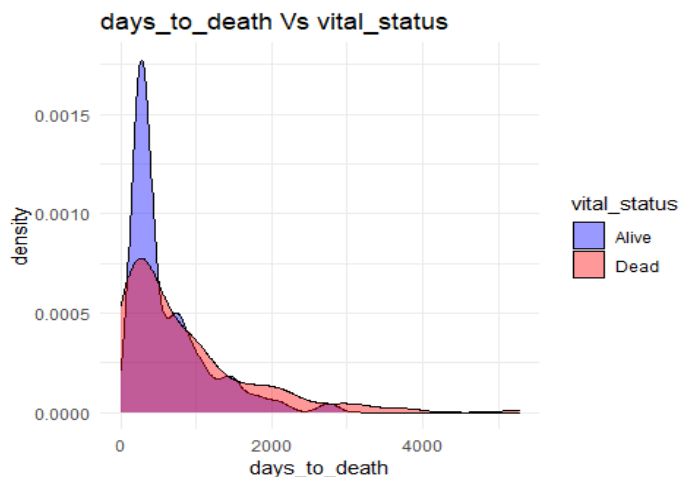


Figure 3.6 Distribution of `days_to_death` concerning levels of `vital_status`

The plot in Figure 3.6 shows the density plots of the sub-samples of Alive and Dead subjects in the data. The density plot of the Alive subsample has high kurtosis as compared to the “dead” subsample of the data. Both the density curves are positively skewed. The density plot of survival time in survival analysis is often positively skewed due to the nature of survival data. Survival data typically

consists of right-censored observations, where the event of interest (such as death or failure) has not occurred for some individuals by the end of the study or observation period. This censoring results in individuals' survival times being known only partially, leading to a truncation of the survival time distribution. As a result, the observed survival times tend to be concentrated on the left side of the distribution, with fewer individuals experiencing the event as time progresses.

This characteristic of survival data contributes to the positively skewed density plot. The density is higher on the left side, representing individuals who experienced the event early on, and gradually tapers off to the right, indicating those who have not yet experienced the event due to censoring. Additionally, there may be other factors that contribute to the skewness, such as heterogeneity in the population, varying risks over time, or different survival dynamics between groups. It is worth noting that the shape of the density plot can be influenced by the specific characteristics of the data and the underlying survival distribution. Therefore, it is important to interpret and analyze survival data considering the context of the study and the specific assumptions of the survival model being used. In this study, three types of survival analysis models have been used namely Kaplan Meier (KM) Model [42], Cox Proportional Hazards Model (CPH) [43], and Accelerated Failure Time (AFT) Model [44]. Each survival model is based on survival and hazard functions. A brief overview of all three models is presented in the following sub-sections.

- **Survival and Hazard Functions**

To describe survival data, two related probabilities are used: the survival probability and the hazard probability. The survivor function $S(t)$, also known as the survival probability, is the probability that an individual survives from the time of origin (e.g., cancer diagnosis) to a specified future time t . The hazard, denoted by $h(t)$, is the probability that an individual being observed at time t will experience an event at that time. In contrast to the survivor function, which focuses on?

the absence of an event, the hazard function focuses on the occurrence of an event.

3.3.1 Kaplan Meier (KM) Model Estimate

The Kaplan-Meier (KM) method is a non-parametric method for estimating the probability of survival based on observed survival times (Kaplan and Meier, 1958) [45]. $S(t_i)$, the survival probability at time t_i , is calculated as follows:

$$S(t_i) = S(t_{i-1})(1 - d_i/n_i) \quad (3.1)$$

Where,

$S(t_{i-1})$ = the probability of being alive at t_{i-1} , n_i = the number of patients alive just before t_i , d_i = the number of events at t_i , $t_0 = 0$, $S(0) = 1$.

The estimated probability ($S(t)$) is a step function that changes value only when an event occurs. It is also possible to compute survival probability confidence intervals. The KM survival curve, which is a plot of the KM survival probability vs. time, provides a useful summary of the data that can be used to estimate measures like median survival time. First, we construct a baseline Kaplan Meier (KM) model that is not adjusted for any group variable. Following that, we modified some variables for survival analysis following the model's format. Gender variable levels were numerically converted. Similarly, changes were made to other variables. Some variables are found to be significant while others are found to be insignificant.

3.3.2 Cox Proportional Hazards (CPH) Model

The Cox proportional hazard model (CPH) is a semi-parametric regression model. The log of hazard ratio is assumed to be linearly dependent on the covariates. The goal of the Cox Proportional Hazard Model is to model the survival rate and event based on a set of features. We considered two types of datasets in the dataset: clinical datasets and image datasets. Kaplan Meier Models have already been explained and used with clinical datasets. There are 64 columns and 504 observations in the image dataset. Each image dataset cell contains either zero or a value greater than zero. These values

represent the pixel values of the 8*8 matrix. The image was converted to an 8X8 matrix, and a vector of 64 values was obtained for each observation by matrix flattening.

Furthermore, two types of features were extracted from the image data: distribution-based features and Gabor-based distribution-based features. The set of distribution features are drawn by row wise operations to calculate mean, standard deviation, sum, maximum value, quartile2, quartile5, quartile10, quartile15, quartile20, quartile25, quartile30, quartile35, quartile40, quartile45, quartile50, quartile55, quartile60, quartile70, quartile75, quartile80, quartile85, quartile90, quartile95, quartile99 and mean absolute deviation. A Gabor filter, named after Dennis Gabor, is a linear filter used in image processing for texture analysis. It analyses whether there is any specific frequency content in the image in specific directions in a localized region around the point or region of analysis. The Gabor filter was used to extract more features from image data, and then distribution-based features were extracted from the Gabor features. The Gabor filter produced two types of data frames: magnitude and energy aptitude datasets. Distribution-based features were created using both datasets. 78 features were used in the resulting dataset to build CPH models.

The Cox proportional-hazards model [46] is essentially a regression model that is commonly used in medical research to investigate the relationship between patient survival time and one or more predictor variables.

- **Basics of the Cox proportional hazards model**

The model's goal is to assess the impact of multiple factors on survival at the same time. In other words, it enables us to investigate how specific factors influence the rate of occurrence of a specific event (e.g., infection, death) at a given point in time. The hazard rate is another name for this rate. In the survival-analysis literature, predictor variables (or factors) are commonly referred to as covariates. The hazard function, denoted by h , expresses the Cox model (t). In a nutshell, the hazard function represents the risk of dying at time t . It can be calculated as follows:

$$h(t)=h_0(t)\times\exp(b_1x_1+b_2x_2+\dots+b_px_p) \quad (3.2)$$

where t denotes the survival time. The hazard function ($h(t)$) is determined by a set of p covariates (x_1, x_2, \dots, x_p), and the coefficients (b_1, b_2, \dots, b_p) measure the impact (i.e., the effect size) of covariates. The baseline hazard is defined as h_0 . It corresponds to the hazard value when all x_i is equal to zero (the quantity $\exp(0)$ equals 1). The 't' in $h(t)$ indicates that the hazard may change over time.

The Cox model can be expressed as a multiple linear regression of the logarithm of the hazard on the variables x_i , with the baseline hazard acting as a time-varying 'intercept' term. The quantities $\exp(b_i)$ are referred to as hazard ratios (HR). A 'b_i' value greater than zero, or a hazard ratio greater than one, indicates that as the value of the i th covariate increases, so does the event hazard and thus the length of survival.

In other words, a hazard ratio greater than one indicates a covariate that is positively associated with event probability and thus negatively associated with survival length.

In summary, HR = 1, No effect HR < 1, Reduction in the hazard HR > 1, Increase in Hazard.

Three Cox Proportional Hazards Model (CPH Model) variants were created: the CPH Model, CPH Model clinical, and CPH Model image. Both clinical and image data were used in the CPH Model. To estimate the hazard function, only clinical data was used in the CPH Model Clinical and image data was used in the CPH Model image.

3.3.3 Accelerated Failure Time (AFT) Model

The accelerated failure time model, like the generalized linear model (GLM), is an extension of the standard linear model that accounts for non-linearity and specific types of data. AFTs are an important class of models because they can handle censored, highly skewed data, which is exactly what one would expect to collect when analyzing machine failure times or the survival times of a group of patients under study.

Because AFTs [47] are fully parametric models, they are unique in the field of survival/reliability analysis. This gives the ability to make inferences like estimating tail probabilities that would be difficult in a non- or semi-parametric framework. The trade-off is that a specific survival time

distribution must be assumed, which may be incorrect. The following is the structure of accelerated failure time models. Given a data matrix X , we observe a vector of survival times (failure times, in the reliability literature) T . It is assumed that the covariates of X have a linear effect on the log of the survival times. Because T is non-negative, we model the effect of the linear predictor X_β on $\log(T)$. The model is

$$F(T|X) = F(\log(T) - X_\beta \sigma) \quad (3.3)$$

F denotes a vectorized, standard distribution function; X_β is called the linear predictor; and σ is called the scale parameter. The survivor function $S(T|X)$ is the probability that a unit will fail after time T . $S(T|X) = 1 - F(T|X)$. As a result, the AFT model belongs to the log-linear model family. The standard normal, standard logistic, and standard smallest extreme value distribution functions are all examples of $F(\cdot)$. We can more clearly write the model as

$$\log(T) = X_\beta + \sigma \varepsilon \quad (3.4)$$

where $\varepsilon \sim F$ to make the linear effect of β on $\log(T)$ a bit more apparent.

Survival models, like generalized linear models, are fitted using a maximum likelihood procedure. This is especially useful because it allows a practitioner to include censored data in the statistical model. The fact that AFTs are fully parametric and can account for data censoring was the primary motivation for including them in ciTools. We assume that AFTs are fitted in R using the survival library's `survreg` function. Numerous AFT models can be used to model survival time and survival status, including log-normal, exponential, Weibull's, and log-logistic models.

3.4 Classification

One of the main reasons for the growing rate of lung cancer is due to lack of accurate and early prediction. Rather than just relying on statistical methods, it is necessary to analyze the data and make proper predictions using automated learning approaches. After performing a statistical evaluation to determine the significance of different variables using the explained statistical models,

this research work is extended to perform prediction of cancer at early stages. Therefore, the classification task is performed using a robust, accurate, and more effective learning approach. Deep learning is a subset of machine learning that is primarily used for Artificial Intelligence (AI) tasks. AI is a field in which computers are trained to perform tasks that humans can perform. The analysis obtained using the statistical models reveals the robustness of image data over clinical data as it has fewer chances of human errors. Therefore, this study utilized image data as an input to the classifier in this phase. The main purpose of applying the classification approach is to identify an image of the Lung being affected by cancer, by developing a deep learning model i.e. multilayer CNN with high performance. The presented CNN is specially designed to enable the most accurate diagnosis of cancer that is superior to other traditional statistics. It can provide accurate results without requiring any type of explicit programming. Further, the result of the developed model is compared with another two algorithms including Random Forest (RF) and XgBoost to evaluate their performance and effectiveness. A brief overview of all proposed and other models is presented as under:

3.4.1 Multilayer CNN

A Convolutional Neural Network (ConvNet/CNN) is a Deep Learning algorithm that can take an input image, assign importance (learnable weights and biases) to different aspects/objects in the image, and distinguish one from the other. Convolutional networks were inspired by biological processes[48], as the pattern of connectivity between neurons resembles the organization of the animal visual cortex. Individual cortical neurons only respond to stimuli in a narrow region of the visual field known as the receptive field. Different neurons' receptive fields partially overlap, covering the entire visual field. When compared to other image classification algorithms, CNNs require very little pre-processing. This means that, in contrast to traditional algorithms, the network learns to optimize the filters (or kernels) through automated learning. This freedom from prior knowledge and human intervention in feature extraction is a significant benefit.

Convolutional neural networks differ from other neural networks in that they perform better with image, speech, or audio signal inputs. They have three different types of layers. A convolutional network's first layer is the convolutional layer. While convolutional layers can be followed by more convolutional layers or pooling layers, the final layer is the fully-connected layer. The CNN becomes more complex with each layer, identifying larger portions of the image. Earlier layers concentrate on simple features like colors and edges. As the image data progresses through the CNN layers, it begins to recognize larger elements or shapes of the object, eventually identifying the intended object [49].

- **Dataset Split**

In the lungs, there are three kinds of cancers i.e. Adenocarcinoma, Large Carcinoma, and Squamous Carcinoma. The data set is a collection of lung CT scan images of these three types of cancers and without cancer (normal) CT scan images. The images in the data set are in PNG format and a few images are also in JPEG format. In Deep Learning, the data set is divided into training and validation sets. Data from the Training set is used to train the deep learning model and a valid set is used to evaluate the deep learning model to predict the status from the images and the predicted output is compared with the actual status of the image. The data is divided into training and validation sets when building a machine-learning model. The machine learning model is trained using the training set. In this case, we're constructing a deep learning model. A sample of images used in this survival analysis is presented in Figure 3.7. In the upper row, the CT scan images are normal and in the lower row, the CT scan images are lung cancer affected.

The structure for extracting images to train the model is as follows in Table 3.3. To carry out the experimentation, a search is performed to determine the required publicly available dataset which can be considered for the classification purpose. The searched dataset [50] is used directly which comprises of different samples comprising of cancer and normal images. Deep learning is a sequential learning model that connects multiple layers of mathematical models with activation

functions. The information from an image's matrices is transitioned from one layer to the next until it reaches the final layer, the output layer.

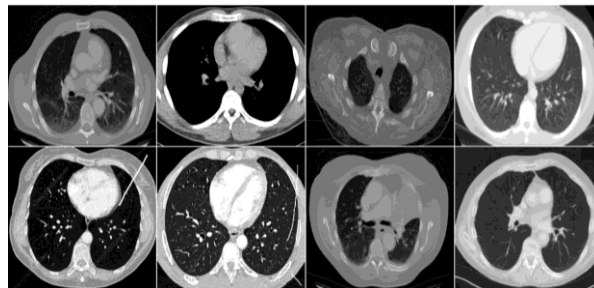


Figure 3.7 A sample of the used image dataset [50]

Table 3.3 Dataset Statistics

Total training normal images	132
Total training cancer images	465
Total testing of normal images	85
Total testing cancer images	322

- **CNN Model Architecture**

CNN model architecture is made of a sequence of layers that pass useful information from images to the layers and from layers to the output layer as shown in Figure 3.8. The summary of the CNN model is shown in Table 3.4.

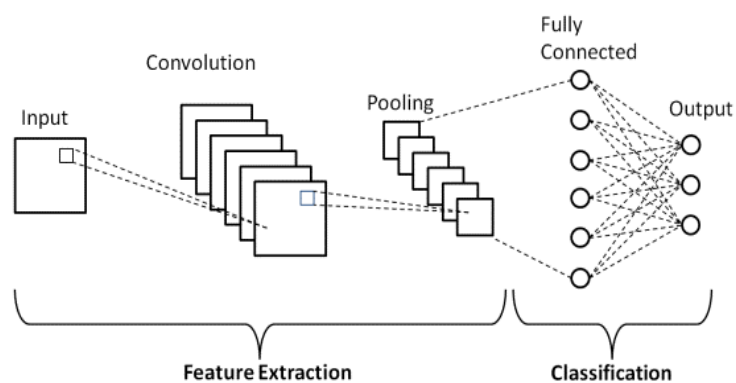


Figure 3.8 The basic architecture of CNN [50]

Table 3.4 CNN Architecture used in this work

Model: "sequential"

Layer (type)	Output Shape	Param #
conv2d_2 (Conv2D)	(None, 148, 148, 16)	448
max_pooling2d_2 (MaxPooling2D)	(None, 74, 74, 16)	0
conv2d_1 (Conv2D)	(None, 72, 72, 32)	4640
max_pooling2d_1 (MaxPooling2D)	(None, 36, 36, 32)	0
conv2d (Conv2D)	(None, 34, 34, 64)	18496
max_pooling2d (MaxPooling2D)	(None, 17, 17, 64)	0
flatten (Flatten)	(None, 18496)	0
dense_1 (Dense)	(None, 512)	9470464
dense (Dense)	(None, 1)	513

=====
Total params: 9,494,561
Trainable params: 9,494,561
Non-trainable params: 0

The activation functions between the layers work as nodes. From these nodes, the information transitions from one layer to the next layer. We use the Library “keras” which is extensively used for deep learning tasks. The convolutional to flatten layers are responsible for performing feature extraction while fully-connected layer performs the classification task. A brief description of the different layers in CNN [51] is given below:

(1) Convolutional Layer

A Convolutional Neural Network's first layer is always a Convolutional Layer. Convolutional layers perform a convolution operation on the input and pass the result to the following layer. Convolutions combine all of the pixels in their receptive field into a single value. For example, if you apply convolution to an image, you will reduce the image size while also combining all of the information in the field into a single pixel. The convolutional layer's final output is a vector.

The 2D convolution layer is the most commonly used type of convolution and is usually abbreviated as conv2D. In a conv2D layer, a filter or kernel "slides" over the 2D input data, performing element-wise multiplication. As a result, the results will be summed into a single output pixel. Therefore, to extract features from input samples, Convolution layer is used. The set of filters in this layer primarily helps in performing the task of feature extraction. As the complexity of CNN increases,

the convolutional layers captures complex features from image. The convolution of a portion of considered data sample is taken to extract the relevant features.

There are three layers in the model. ReLU, an activation function, removes negative values from all matrices and replaces them with zero while leaving positive values alone. This function can be used in conjunction with convolution layers. The convolution layers are positioned at the first, third, and fifth layers.

(2) Max-Pooling Layer

The feature maps' dimensions are reduced by pooling layers. As a result, it reduces the number of parameters to learn as well as the amount of computation done in the network. The pooling layer summarizes the features in a region of the feature map produced by a convolution layer. As a result, subsequent operations are carried out on summarized features rather than precisely positioned features generated by the convolution layer. As a result, the model is more resistant to changes in the position of the features in the input image. These layers downsample the feature map to introduce Translation invariance, which reduces the CNN model's overfitting.

- **Max Pooling Process**

Pooling that selects the maximum element from the region of the feature map covered by the filter is known as max pooling. As a result, the output of the max-pooling layer would be a feature map that contained the most prominent features of the previous feature map. Applying the MaxPool2D layer to the matrix yields the MaxPooled output in tensor form. When applied to the matrix, the Max pooling layer will traverse the matrix, computing the maximum of each 2x2 pool with a jump of 2. Print the tensor's shape. To remove dimensions of size 1 from the shape of a tensor, use `tf.squeeze`. In this model, there are three such layers. The max-pooling layers are positioned at the second, fourth, and sixth layers.

(3) Flatten Layer

Flattening is the next step in the process. Flattening is the process of combining all of the resulting 2-Dimensional arrays from pooled feature maps into a single long continuous linear vector. To classify the image, the flattened matrix is fed as input to the fully connected layer. In this model, one flattened layer is used to fulfill the purpose.

(4) Fully-connected (Dense Layer)

A neural network's fully-connected layer, also known as the dense layer, is always the last. As is the case with feed-forward neural networks, neurons in fully connected layers are fully connected to all activations in the previous layer. Before applying the softmax classifier, it is common to use one or two fully connected layers. The fully connected operates with a flattened input, which means that each input is coupled to every neuron. The flattened vector is then sent through a few more fully connected layers, where the mathematical functional operations are normally performed. At this point, the classification procedure begins. If FC layers are present, they are usually found near the end of CNN architectures. In the present model, two such layers are added. The final layer of this type is an output layer with a softmax activation function. The dense layer functions as an NN hidden layer, with 512 hidden neurons in a ReLU unit, and operates on input to provide output in the form of a predicted class label.

The CNN model is fine-tuned with optimizer rmsprop at a 0.001 learning rate. The data must be transformed and scaled to meet the standards of the CNN model. To prepare data for the CNN model, we use the `image_data_generator` command. After transformation, it declares the number of inputs and the number of outputs from the inputs. The CNN model in this project has two outcomes/classes: normal CT scan and cancer-affected CT scan. The CNN model will assign one of the two classes to the CT scan image.

The model takes a long time to compile because CNN is computationally expensive. It requires enough RAM and processing power to run smoothly and quickly. For the task, we used 4GB RAM and an Intel Processor 5th Generation HP Laptop.

3.4.2 Random Forest (RF)

Random Forest is a versatile and powerful supervised machine learning algorithm that grows and combines multiple decision trees to form a "forest." It applies to classification and regression problems. Another type of algorithm used to classify data is the decision tree. In the most basic sense, it is a flowchart that draws a clear pathway to a decision or outcome; it begins at a single point and then branches off into two or more directions, with each branch of the decision tree offering different possible outcomes. The ensemble of trees produced by Random Forest returns either the mode or mean of the individual trees. This method relies on a large number of trees to produce more accurate and stable results [52].

The Random Forest model is based on the idea that multiple uncorrelated models (individual decision trees) perform much better as a group than they do individually. When Random Forest is used for classification, each tree provides a classification or "vote." The classification with the most "votes" is chosen by the forest. When using a Random Forest for regression, the forest chooses the average of all tree outputs. The key point here is that there is little (or no) correlation between the individual models—that is, the decision trees that comprise the larger Random Forest model. While individual decision trees may make mistakes, the majority of the group will be correct, moving the overall result in the right direction.

The "bagging" method is commonly used to train decision trees in an ensemble, as well as trees in a Random Forest. The "bagging" method is a Bootstrap Aggregation ensemble machine learning algorithm. An ensemble method uses predictions from multiple machine learning algorithms to make more accurate predictions than a single model. Random Forest is an ensemble method as well. When analyzing a large database, a random forest is far more efficient than a single decision tree. Random

Forest, on the other hand, is less efficient than a neural network. Neural nets are more complex than random forests, but they produce the best results possible by adapting to changing inputs [53].

3.4.3 XgBoost Model

The XgBoost is an acronym for Extreme Gradient Boosting classifier is a machine-learning algorithm for structured and tabular data. XgBoost is a gradient-boosted decision tree implementation optimized for speed and performance. In gradient boosting, each predictor corrects the error of its predecessor. Unlike Adaboost, the weights of the training instances are not changed; instead, each predictor is trained using the predecessor's residual errors as labels. This algorithm sequentially generates decision trees. Weights are very important in XgBoost. All of the independent variables are given weights, which are then fed into the decision tree, which predicts results [54]. These individual classifiers/predictors are then combined to form a more powerful and precise model. It can solve problems involving regression, classification, ranking, and user-defined prediction.

XgBoost is a method of ensemble learning. It is not always sufficient to rely on the results of a single machine-learning model. Ensemble learning provides a methodical approach to combining the predictive power of multiple learners. The result is a single model that aggregates the output of several models. The models that comprise the ensemble, also known as base learners, could be from the same or different learning algorithms. Bagging and boosting are two popular ensemble learning methods.

3.5 Performance Evaluation and Output

This forms the last stage in the entire process for predictive analysis of lung cancer. The most well-known Machine Learning model validation method used in evaluating classification problems is accuracy. One reason for its popularity is its ease of use. It is simple to comprehend and implement. In simple cases, accuracy is a good metric to use to evaluate model performance. In this work, accuracy is used as the metric for the performance evaluation of different models.

The confusion Matrix [55] is a performance metric for machine learning classification problems with two or more classes as output. It is a table with four different predicted and actual values as shown in Table 3.5.

Table 3.5 Confusion Matrix

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

- **Terminology**

- (1) True positives (TP): These are cases in which we predicted yes (they have the disease) and they do.
- (2) True negatives (TN): We predicted no, and they are disease-free.
- (3) False positives (FP): Although we predicted yes, they do not have the disease. (This is also referred to as a "Type I error.")
- (4) False negatives (FN): We predicted that they would not have the disease, but they do. (This is also referred to as a "Type II error.")

In machine learning, the accuracy score is an evaluation metric that compares the number of correct predictions made by a model to the total number of predictions made. We compute it by dividing the number of correct predictions by the total number of predictions. Mathematically, it can be defined as

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \tag{3.5}$$

Further, the loss is also calculated. A loss is a penalty for making an incorrect prediction. In other words, the loss is a number that indicates how inaccurate the model's prediction was on a single example. The loss is zero if the model's prediction is perfect; otherwise, the loss is greater. Finally, the classification result is obtained by predicting the lung image with cancer and non-cancer.

In addition to accuracy, other performance evaluation metrics are also available such as precision, recall, etc. Precision refers to the number of actual positives out of the projected positive values and can be defined as:

$$\text{Precision} = \frac{TP}{TP+FP} \quad (3.6)$$

By classifying anything as Positive (True Positive), Recall estimates the number of Actual Positives that the model captures which can be defined as.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3.7)$$

If there is an uneven class distribution (many Actual Negatives) and we need to identify a balance between Precision and Recall, the F1 Score might be a better metric to choose. F1-score is the harmonic mean of precision and recall.

$$F1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (3.8)$$

Chapter 4

Results and Discussion

This chapter provides the results for pre-processing, association analysis, and statistical as well as deep learning methods. Firstly, the results obtained for missing values treatment and association of variables are provided. Next, the results for three statistical models i.e. KM model, CPH model, and AFT model are provided which were employed to perform the survival risk analysis of lung cancer patients. The output achieved reveals the most significant variables having a crucial effect on patients using p-value and the data which is proved to be robust in such analysis. Further, the result achieved by applying an efficient deep learning (DL) model is compared with the other two ensemble techniques to evaluate their performance in identifying the images with lung cancer and without lung cancer based on accuracy rate. Thus, the analysis obtained from this study can be helpful to researchers and clinicians in performing a risk analysis of lung cancer more efficiently.

4.1 Missing Values Treatment Results

After MissForest imputation through the random forest algorithm, the sanity of the data was checked and 0% missing values in the data were found. The distribution of the days_to_death variable is also checked through histogram in data1 and data_impute data frames. If the distribution of the two datasets is near to each other, we can say both datasets come from the same population. Figure 4.1 shows the distribution of the days_to_death variable before imputation and after imputation. Statistical test for comparing means of days_to_death variable in data1 and data_impute data frames is as under:

The output of T test:

Test	Value
statistic.t	-2.0356

p.value	0.043
estimates. mean of x	679.0426
estimates. mean of y	835.5987
stderr value	76.9075

The t.test result supports the argument that the means of days_to_death variables is the same in both data1 and data_impute data frames. We also checked the difference in the variance of the days_to_death variable of both datasets.

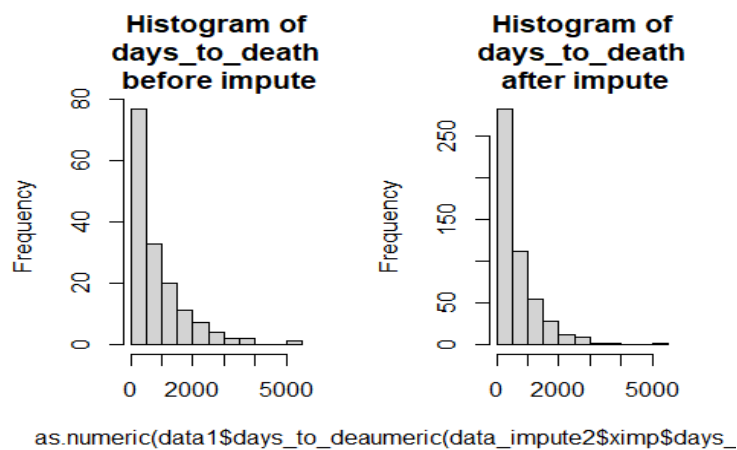


Figure 4.1 Comparing Distribution of Datasets

The following data shows the result of the F test at a 0.05 significance level:

Test	Value
statistic. F	1.7717
parameter. num df	156
parameter. denom df	503
p.value	0
conf.int1	1.3852
conf.int2	2.3056

estimate.ratio of variances value 1.7717

Variance has almost doubled in days_to_death variables from data1 and data_impute. With the boxplot, we can check the outliers which may be the possible cause of high variance as shown in Figure 4.2.

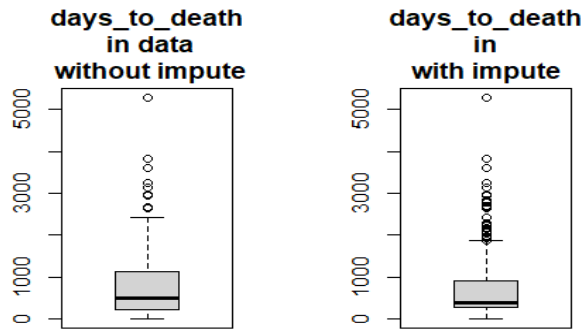


Figure 4.2 Boxplot for outliers

4.2 Results for Association Analysis

In Figure 4.3, we can find a positive correlation of days_to_death with days_to_last_followup while a negative correlation with year_of_pathological_diagnosis. The sign (X) in the corrplot represents that the correlation coefficient is not significant at a 0.05 significance level. Similarly, we conducted a Chi-Squared test of all variables with the vital_status variable to explore the association of vital_status with other variables.

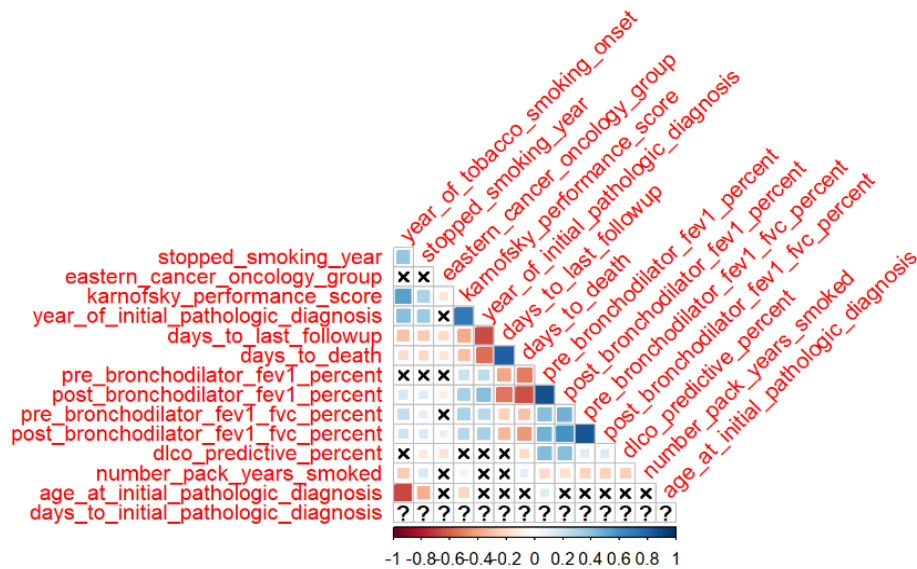


Figure 4.3 Corrplot to represent correlation

If the value of the chi-square is greater than some threshold, it shows an association between non-numeric variables also called factor variables. The association between numeric variables is called correlation i.e. $P = 0.05$. It is further to be noted that continuous scale, this test cannot be applied. Thus, the test of independence is applied to check whether two-factor variables are independent of each other or not. The dependence means that one-factor variable influences the distribution of levels in the other factor variable.

In a chi-square test, the degrees of freedom (df) are determined by the number of categories or groups being compared, as well as any constraints or conditions imposed on the data. The degrees of freedom represent the number of values that are free to vary in the test statistic, given certain constraints. To explain this further, let's consider an example: Suppose we are interested in examining whether there is a relationship between hair color and eye color in a population. We collect data from a sample of individuals and categorize them based on their hair color (blonde, brown, black) and eye color (blue, brown, green). We want to test the null hypothesis that hair color and eye color are independent. To determine the degrees of freedom for the chi-square test, we use the formula: $df = (r - 1) * (c - 1)$, where r is the number of rows (hair colors) and c is the number of columns (eye colors) in the contingency table. In this example, we have $r = 3$ (Blonde, Brown, Black)

and $c = 3$ (Blue, Brown, Green). Thus, the degrees of freedom for the chi-square test would be $df = (3 - 1) * (3 - 1) = 2 * 2 = 4$. In this case, we have four degrees of freedom, indicating that the chi-square test will use the chi-square distribution with four degrees of freedom to evaluate the statistical significance of the observed association between hair color and eye color.

Table 4.1 Chi-squared results for association analysis

	X	Y	Chi.Square	p.value
X-squared	tissue_prospective_collection_indicator	vital_status	83.843	0.000
X-squared1	gender	vital_status	0.006	0.940
X-squared2	ethnicity	vital_status	5.065	0.024
X-squared3	tissue_retrospective_collection_indicator	vital_status	83.843	0.000
X-squared4	race	vital_status	7.999	0.018
X-squared5	other_dx	vital_status	14.224	0.003
X-squared6	anatomic_neoplasm_subdivision	vital_status	6.322	0.503
X-squared7	history_of_neoadjuvant_treatment	vital_status	3.826	0.148
X-squared8	location_in_lung_parenchyma	vital_status	1.912	0.167
X-squared9	histological_type	vital_status	3.952	0.413
X-squared10	residual_tumor	vital_status	2.969	0.396
X-squared11	system_version	vital_status	109.200	0.000
X-squared12	pathologic_T	vital_status	36.162	0.000
X-squared13	pathologic_N	vital_status	9.800	0.044
X-squared14	pathologic_M	vital_status	6.259	0.181
X-squared15	pathologic_stage	vital_status	28.643	0.001
X-squared16	pulmonary_function_test_performed	vital_status	25.285	0.000
X-squared17	kras_gene_analysis_performed	vital_status	0.439	0.507
X-squared18	egfr_mutation_performed	vital_status	0.334	0.563
X-squared19	eml4_alk_translocation_performed	vital_status	0.984	0.321
X-squared20	tobacco_smoking_history	vital_status	4.650	0.460
X-squared21	vital_status	vital_status	499.411	0.000
X-squared22	performance_status_scale_timing	vital_status	4.569	0.206
X-squared23	radiation_therapy	vital_status	2.931	0.231
X-squared24	postoperative_rx_tx	vital_status	19.641	0.000
X-squared25	primary_therapy_outcome_success	vital_status	4.169	0.244
X-squared26	person_neoplasm_cancer_status	vital_status	179.412	0.000
X-squared27	new_tumor_event_after_initial_treatment	vital_status	208.370	0.000

From Table 4.1, the association of the following variables was not found significant with vital_status as the P value is greater than 0.005.

gender, eastern_cancer_oncology_group, anatomic_neoplasm_subdivision, location_in_ling_parenchyma, histological_type, residual_tumor, system_version, pathologic_M days_to_death, pre_bronchodilator_fev1_percent, post_bronchodilator_fev1_percent, pre_bronchodilator_fev1_fvc_percent, post_bronchodilator_fev1_fvc_percent, kras_gene_analysis_performed, dlco_predictive_percent, kras_mutation_found, egfr_mutation_performed, eml4_alk_translocation_performed, tobacco_smoking_history, number_pack_years_smoked, radiation_therapy, primary_therapy_outcome_success, days_to_birth, age_at_initial_pathologic_diagnosis, and duration_of_smoking

The following set of variables was found to have a significant association with vital_status as the P value is less than 0.005.

tissue_prospective_collection_indicator, ethnicity, tissue_retrospective_collection_indicator, year_of_tobacco_smoking_onset, stopped_smoking_year, karnofsky_performance_score, race, other_dx, system_version, pathologic_T, pathologic_N, days_to_last_followup, pathologic_stage, performance_status_scale_timing, postoperative_rx_tx, person_neoplasm_cancer_status, new_tumor_event_after_initial_treatment, anatomic_neoplasm_subdivision_other, and tissue_source_site.

4.3 Statistical Models Results

The implication of different statistical models yielded different results to determine the most relevant variables among all and the reliable data. The results achieved for each model are explained one by one in the subsequent subsections.

4.3.1 Kaplan-Meier (KM) Model

In many researches, multiple models are built through successive improvements. To compare the successive improvements and the models, we compare them with the first acceptable model. The model with which we compare is called BASELINE MODEL. Therefore, First, we construct a baseline Kaplan Meier (KM) model that is not adjusted for any group variable.

- **KM Model without Group Variable**

We modified some variables for survival analysis to fit the model's format. Gender variable levels were converted to numbers. Similarly, modifications were made to other variables.

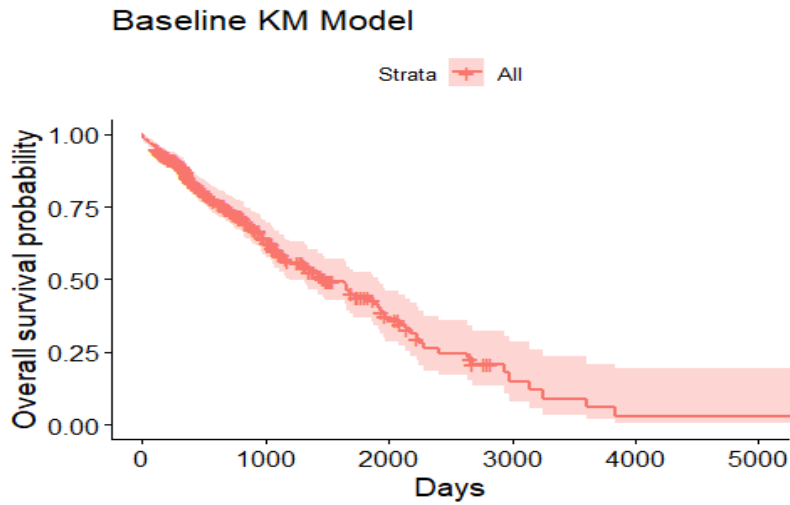


Figure 4.4 Survival curve of the baseline KM model

Strata is the label to the variable which shows the levels in the factor variable but in this chart i.e. given in Figure 4.4, the variable is numeric, not a factor variable hence it shows, by default, All.

The model shows the survival probability vis-a-viz the days_to_death. We can see that the lower the days_to_death, the higher the proportion. Survival probability decreasing as survival time increases is a fundamental characteristic of survival analysis and reflects the underlying concept of time-to-event data. In survival analysis, the survival probability represents the likelihood of an individual surviving beyond a specific time point, given that they have survived up to that time. As time progresses, the survival probability naturally decreases for several reasons:

1. **Occurrence of Events:** As time goes by, events of interest, such as deaths or failures, occur in the population under study. These events lead to a reduction in the number of individuals who have survived up to that point, causing a decline in the survival probability. Each event that occurs decreases the pool of individuals at risk of experiencing the event in the future.
2. **Censoring:** Censoring is a common feature in survival analysis where individuals are not observed until the event of interest occurs or the study ends. Censored observations

contribute to the decrease in survival probability as the study progresses. Censoring implies that the event has not happened for those individuals, and their survival status beyond the observed time is uncertain. This uncertainty contributes to a decline in the estimated survival probability.

3. **Susceptibility to the Event:** Over time, individuals in a population may become more susceptible to the event of interest. Factors such as aging, disease progression, or accumulated risks can increase the likelihood of experiencing the event. As a result, the survival probability decreases as the population becomes more prone to the event with time.

It's important to note that the rate at which the survival probability decreases over time depends on various factors, including the characteristics of the population, the nature of the event, and any interventions or treatments applied. These factors can lead to different shapes and patterns of the survival curve, highlighting the dynamic nature of survival probabilities in different scenarios. By analyzing the survival curve and understanding the decreasing survival probabilities, researchers and clinicians can gain insights into disease progression, treatment efficacy, and prognosis, which are crucial for making informed decisions in medical research and patient care.

- **KM Model with Group Variable**

Next, the baseline KM model was developed by adjusting for grouping variables. A criterion was set to check the significance level of the variables. A p-value less than 0.005 ($p < 0.005$) is considered to be a significant one and vice-versa. On applying the test, some variables were found to be statistically significant while others did not reveal any significance. Figure 4.5 depicts the variables that yielded significant results in this experiment. KM model:

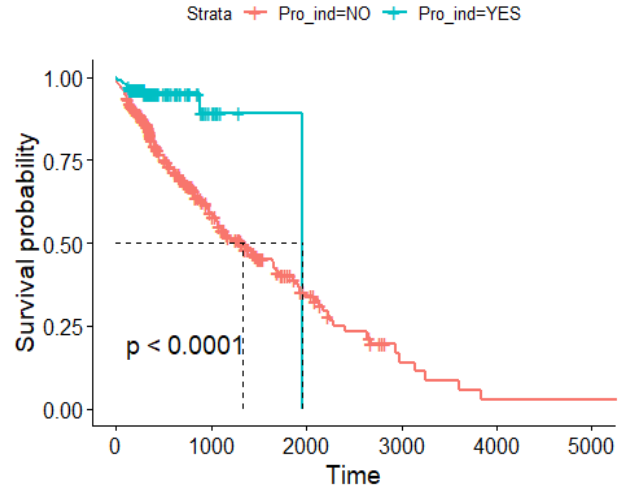
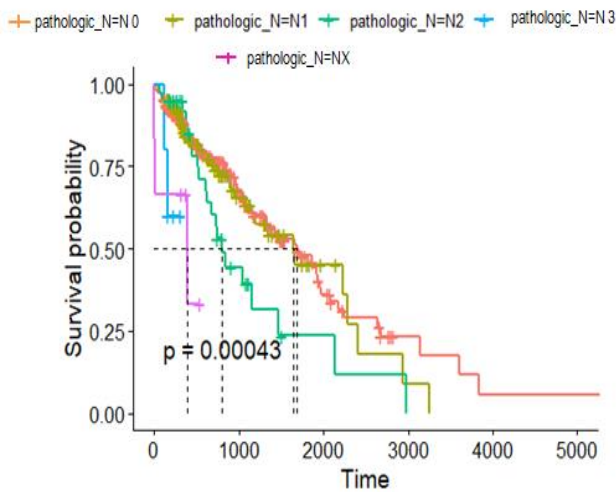
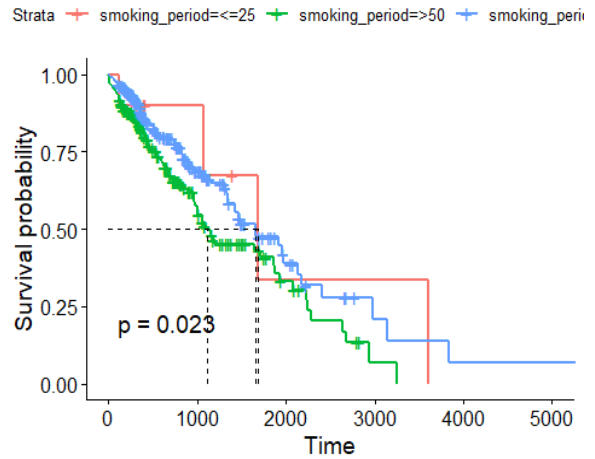
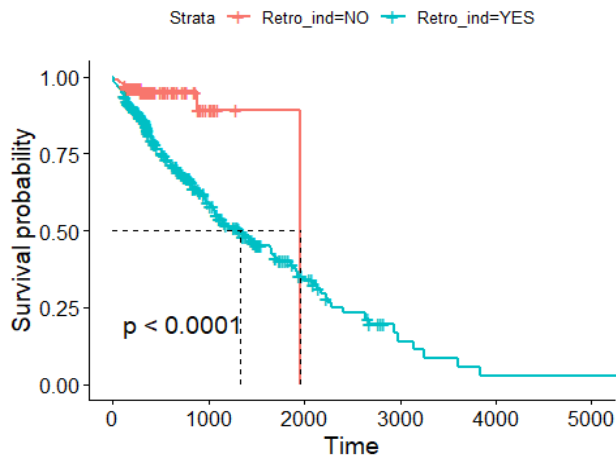
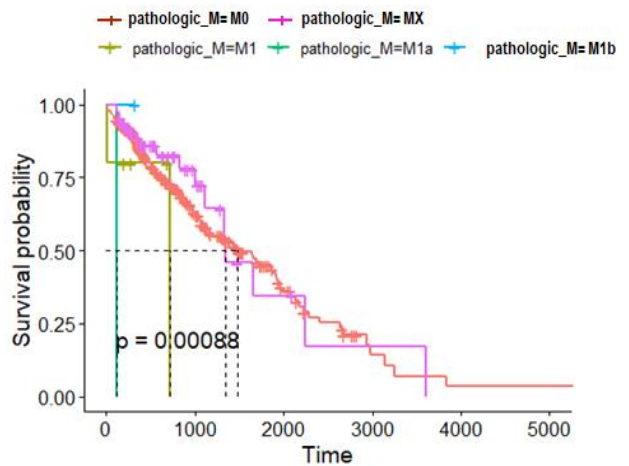
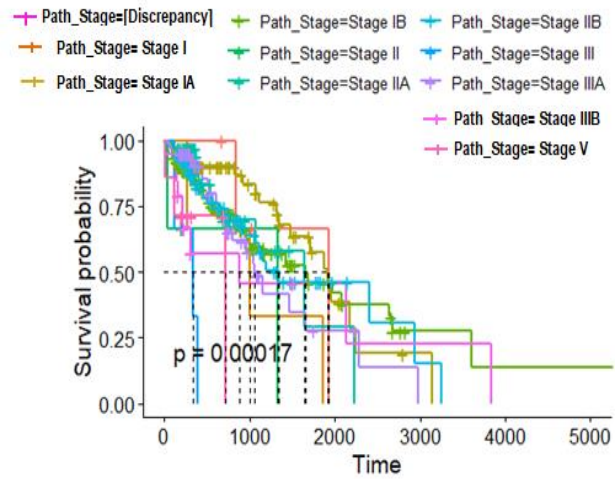


Figure 4.5: (1) with Group Var Pathologic_N (2) with Var tissue_prospective_collection_indicator

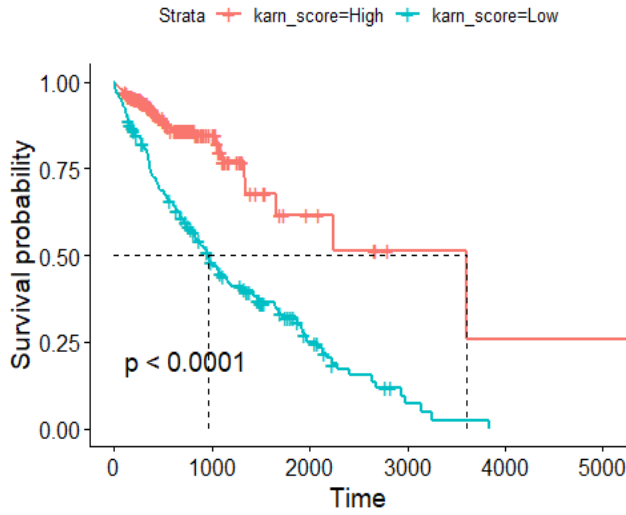


(3) with Group Var tissue_retrospective_collection_indicator (4) with Group Var Smoking Period

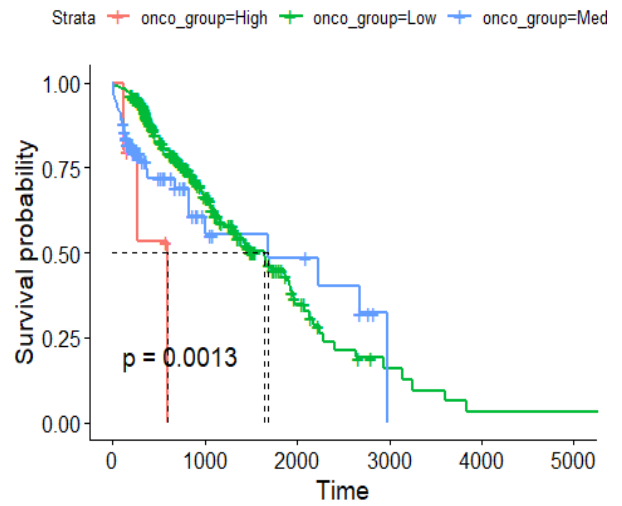


(5) with Group Var pathologic_stage

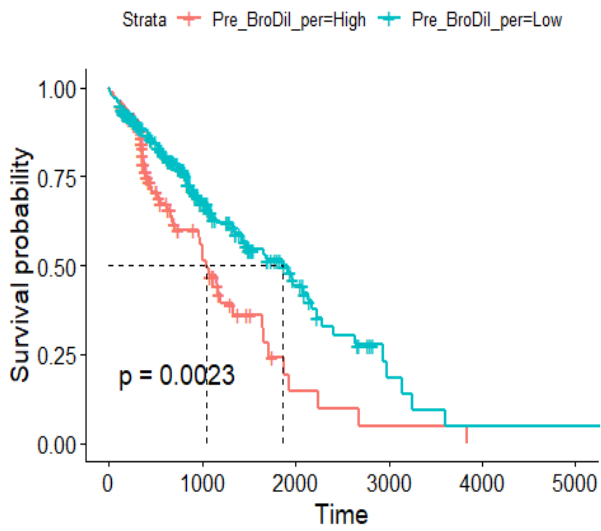
(6) with Group Var pathologic_M



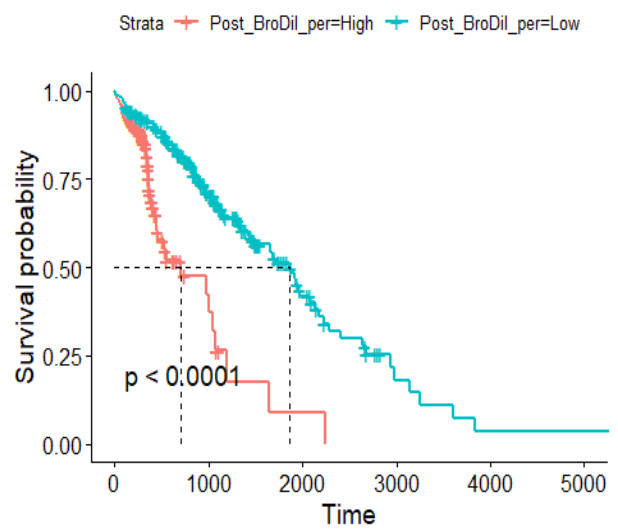
(7) with Var karnofsky_performance_score



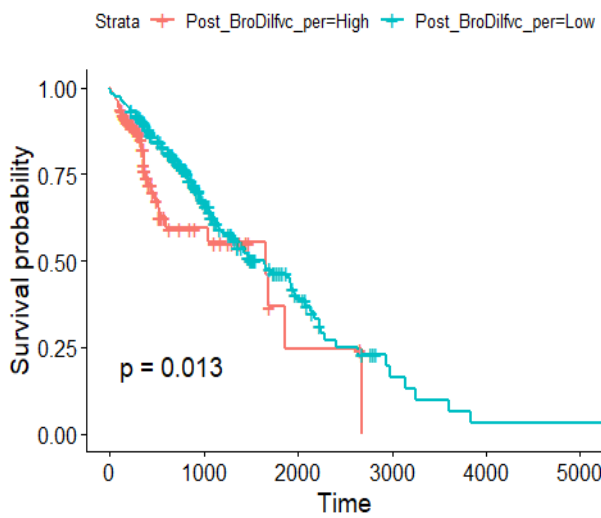
(8) with Group Var eastern_cancer_oncology_group



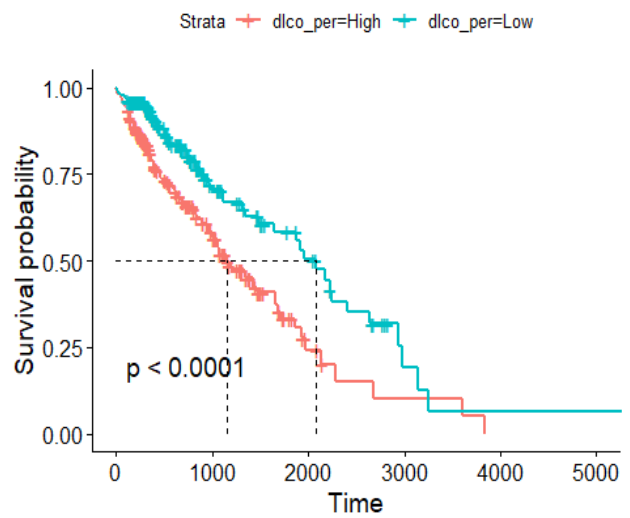
(9) with Var pre_bronchodilator_fev1_percent



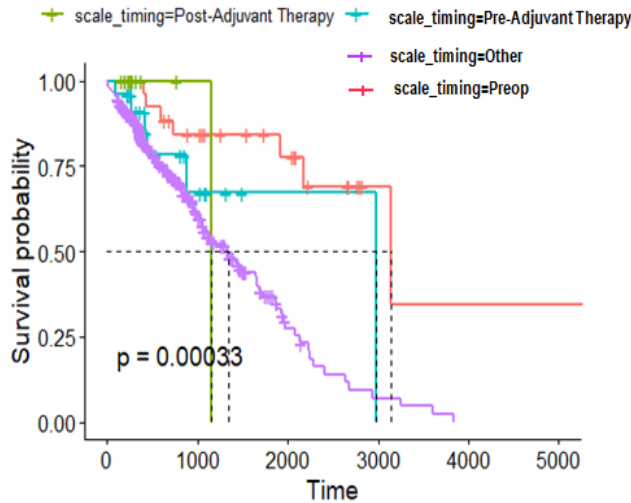
(10) with Var post_bronchodilator_fev1_percent



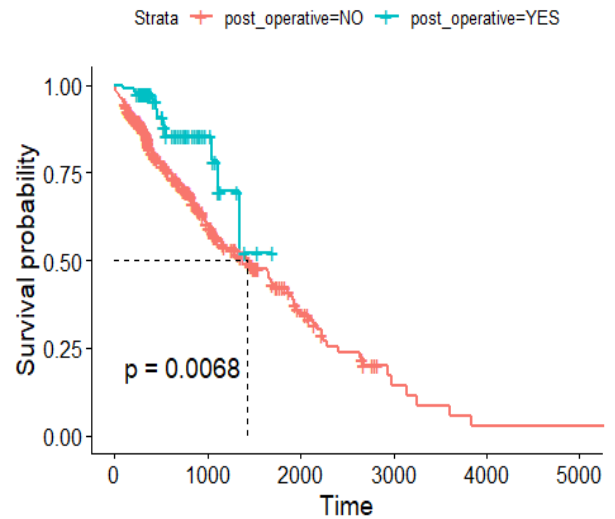
(11) with Var post_bronchodilator_fev1_fvc_percent



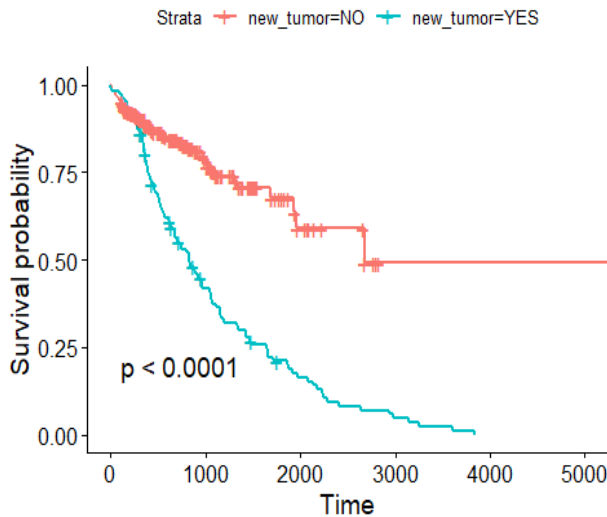
(12) with Var dlco_predictive_percent



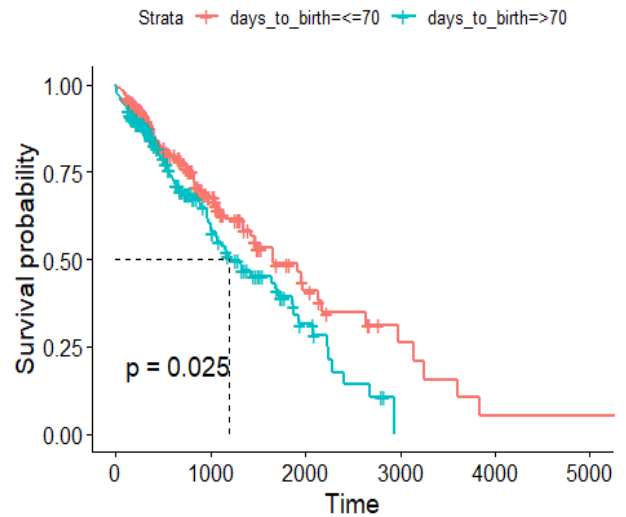
(13) with Group Var performance_status_scale_timing



(14) with Group Var postoperative_rx_tx



(15) with Var new_tumor_event_after_initial_treatment



(16) with Group Var Age_Group

The log-rank test is a commonly used statistical test in survival analysis to assess whether there are significant differences in survival probabilities between two or more groups. It is a nonparametric test that does not assume any specific distribution for the survival times. Here are the main reasons why the log-rank test is performed in survival analysis:

1. **Comparison of Survival Curves:** The log-rank test allows for the comparison of survival curves between different groups or categories. It helps determine whether there are statistically significant differences in survival probabilities among the groups being

compared. This is particularly useful when studying the impact of different treatments, interventions, or patient characteristics on survival outcomes.

2. **Handling Censored Data:** The log-rank test takes into account censored data, which are observations where the event of interest has not occurred by the end of the study period or when individuals are lost to follow-up. Censored observations are included in the analysis and contribute to the estimation of survival probabilities. The test appropriately accounts for censored data and adjusts the comparisons accordingly.
3. **Time-to-Event Data:** Survival analysis often deals with time-to-event data, where the focus is on the time until an event occurs (e.g., death, failure, recurrence). The log-rank test is specifically designed for analyzing such data, allowing for the assessment of differences in survival probabilities over time between groups.
4. **Nonparametric Approach:** The log-rank test is a nonparametric test, which means it does not require assumptions about the underlying distribution of survival times. This flexibility makes it applicable in a wide range of scenarios, including when the survival times do not follow a specific distribution or when the sample size is small.
5. **Hypothesis Testing:** The log-rank test provides a statistical hypothesis test to evaluate whether the observed differences in survival probabilities between groups are statistically significant or due to random chance. It yields a test statistic and a p-value that can be used to make conclusions about the presence or absence of a statistically significant difference in survival.

Overall, the log-rank test is performed in survival analysis to compare survival probabilities between groups, account for censored data, handle time-to-event data, and provide a statistical inference for assessing the significance of differences in survival outcomes. It is a widely used tool in the field and helps researchers and clinicians understand the impact of various factors on survival. As shown in Figure 4.5, due to unique features, in this work, a log-rank test was run to determine if: (1) there

were differences in the survival distribution of lung cancer patients for the pathologic_N types: N0, N1, N2, N3, and NX. The survival distributions for pathologic types were statistically significantly different, $\chi^2(4) = 18.3$ (where 4 is the degree of freedom), $p < 0.05$; (2) there were differences in the survival distribution of lung cancer patients for the tissue_prospective_collection_indicator types: Yes and No. The survival distributions for tissue_prospective_collection_indicator types were statistically significantly different, $\chi^2(1) = 16.7$ (where 1 is the degree of freedom), $p < 0.05$; (3) there were differences in the survival distribution of lung cancer patients for the tissue_retrospective_collection_indicator types: Yes and No. The survival distributions for tissue_retrospective_collection_indicator types were statistically significantly different, $\chi^2(1) = 16.7$ (where 1 is the degree of freedom), $p < 0.05$; (4) there were differences in the survival distribution of lung cancer patients for the smoking_period types: ≤ 25 , 26-50, and > 50 . The survival distributions for smoking_period types were statistically significantly different, $\chi^2(2) = 6.5$ (where 2 is the degree of freedom), $p < 0.05$; (5) there were differences in the survival distribution of lung cancer patients for the pathologic_stage types: Stage I, Stage IA, Stage II, Stage IIA, Stage IB, Stage IIB, Stage III, Stage IIIA, Stage IIIB and Stage IV, The survival distributions for pathologic_stage types were statistically significantly different, $\chi^2(10) = 34.1$ (where 10 is the degree of freedom), $p < 0.05$; (6) there were differences in the survival distribution of lung cancer patients for the pathologic_M types: M0 M1 M1a M1b MX. The survival distributions for pathologic_M types were statistically significantly different, $\chi^2(4) = 18.9$ (where 4 is the degree of freedom), $p < 0.05$; (7) there were differences in the survival distribution of lung cancer patients for the karn_sc ore types: Low (≤ 50) and High (> 50). The survival distributions for karn_score types were statistically significantly different, $\chi^2(1) = 48.8$ (where 1 is the degree of freedom), $p < 0.05$; (8) there were differences in the survival distribution of lung cancer patients for the onco_group types: Low, Medium, and High. The survival distributions for onco_group types were statistically significantly different, $\chi^2(2) = 13.3$ (where 2 is the degree of freedom), $p < 0.05$; (9) there were differences in the

survival distribution of lung cancer patients for the Pre_BroDil_per types. The survival distributions for Pre_BroDil_per types were statistically significantly different, $\chi^2(1) = 9.3$ (where 1 is the degree of freedom), $p < 0.05$; (10) There were differences in the survival distribution of lung cancer patients for the Post_BroDil_per types. The survival distributions for Post_BroDil_per types were statistically significantly different, $\chi^2(1) = 32.9$ (where 1 is the degree of freedom), $p < 0.05$; (11) There were differences in the survival distribution of lung cancer patients for the Post_BroDilfvc_per types. The survival distributions for Post_BroDilfvc_per types were statistically significantly different, $\chi^2(1) = 6.2$ (where 1 is the degree of freedom), $p < 0.05$; (12) There were differences in the survival distribution of lung cancer patients for the dlco_per types. The survival distributions for dlco_per types were statistically significantly different, $\chi^2(1) = 16.2$ (where 1 is the degree of freedom), $p < 0.05$; (13) There were differences in the survival distribution of lung cancer patients for the scale_timing types. The survival distributions for scale_timing types were statistically significantly different, $\chi^2(3) = 18.6$ (where 3 is the degree of freedom), $p < 0.0$; (14) There were differences in the survival distribution of lung cancer patients for the post_operative types. The survival distributions for post_operative types were statistically significantly different, $\chi^2(1) = 7.3$ (where 1 is the degree of freedom), $p < 0.05$; (15) There were differences in the survival distribution of lung cancer patients for the new_tumor types. The survival distributions for new_tumor types were statistically significantly different, $\chi^2(1) = 52.2$ (where 1 is the degree of freedom), $p < 0.05$; (16) There were differences in the survival distribution of lung cancer patients for the days_to_birth groups. The survival distributions for days_to_birth groups were statistically significantly different, $\chi^2(1) = 5.1$ (where 1 is the degree of freedom), $p < 0.05$. The summary of the results achieved in the KM model is presented in Table 4.2. Further, the plots of the hazard ratio and beta coefficients of the KM model with each clinical feature are given in Figures 4.6 and 4.7.

Table 4.2 KM model results

Variable	beta	HR	HR.Lower	HR.Upper	wald. test	p.value
tobacco_smoking_history	16.0000	16.000	0.000	0.000	0.000	0.0e+00
tissue_retrospective_collection_indicator	1.3000	3.800	1.900	7.500	14.000	1.4e-04
new_tumor_event_after_initial_treatment	1.1000	3.200	2.300	4.400	48.000	0.0e+00
pathologic_M	0.9300	2.900	0.620	2.600	0.000	5.2e-01
Ethnicity	0.6500	1.900	0.840	4.400	2.400	1.2e-01
eastern_cancer_oncology_group	0.3100	1.400	1.000	1.800	4.100	4.4e-02
pulmonary_function_test_performed	0.3600	1.400	0.870	2.400	2.000	1.6e-01
days_to_birth	0.3600	1.400	1.000	2.000	5.000	2.5e-02
location_in_lung_parenchyma	0.2500	1.300	0.930	1.800	2.300	1.3e-01
Race	0.1800	1.200	0.940	1.500	2.100	1.5e-01
Gender	0.0780	1.100	0.760	1.500	0.180	6.7e-01
performance_status_scale_timing	0.7500	1.100	0.260	0.960	2.100	1.7e+01
days_to_last_followup	-0.0012	1.000	1.000	1.000	63.000	0.0e+00
pre_bronchodilator_fev1_percent	0.0280	1.000	1.000	1.000	34.000	0.0e+00
post_bronchodilator_fev1_percent	0.0450	1.000	1.000	1.100	60.000	0.0e+00
pre_bronchodilator_fev1_fvc_percent	0.0029	1.000	0.990	1.000	0.170	6.8e-01
post_bronchodilator_fev1_fvc_percent	0.0330	1.000	1.000	1.100	13.000	2.9e-04
dlco_predictive_percent	0.0230	1.000	1.000	1.000	25.000	7.0e-07
number_pack_years_smoked	-0.0028	1.000	0.990	1.000	1.200	2.8e-01
age_at_initial_pathologic_diagnosis	0.0330	1.000	1.000	1.100	10.000	1.4e-03
stopped_smoking_year	-0.0110	0.990	0.970	1.000	1.700	1.9e-01
karnofsky_performance_score	-0.0230	0.980	0.970	0.980	60.000	0.0e+00
year_of_tobacco_smoking_onset	-0.0270	0.970	0.960	0.990	10.000	1.4e-03
egfr_mutation_performed	-0.0300	0.970	0.400	2.400	0.000	9.5e-01
year_of_initial_pathologic_diagnosis	-0.0400	0.960	0.920	1.000	3.600	5.6e-02
anatomic_neoplasm_subdivision	0.5300	0.720	0.150	0.280	0.360	3.6e-01
kras_gene_analysis_performed	-0.3400	0.710	0.180	2.900	0.230	6.3e-01

Variable	beta	HR	HR.Lower	HR.Upper	wald.test	p.value
pathologic_N	0.0590	0.590	0.730	1.100	1.400	1.6e+00
eml4_alk_translocation_performed	-0.5600	0.570	0.080	4.100	0.310	5.8e-01
person_neoplasm_cancer_status	-0.5800	0.560	0.077	0.240	4.100	1.3e+01
histological_type	-0.7500	0.470	0.190	1.200	2.600	1.1e-01
postoperative_rx_tx	-0.8700	0.420	0.220	0.800	6.900	8.7e-03
tissue_prospective_collection_indicator	-1.3000	0.260	0.130	0.520	14.000	1.4e-04
pathologic_T	0.1200	0.014	0.380	0.400	0.870	6.4e-01
other_dx	0.3800	-0.180	0.940	0.310	0.140	2.3e+00
pathologic_stage	0.7900	-0.350	0.370	0.160	0.240	4.1e-01
system_version	19.0000	-3.200	0.000	0.009	0.031	1.2e-02
residual_tumor	1.1000	-14.000	1.300	0.000	0.670	7.6e+00

Hazard ratio (HR): is a measure of the effect of an intervention on an outcome of interest over time. According to convention when Hazard Ratio (HR): HR=1, the variables do not have any significance in the survival probability. When the Hazard Ratio <1 it reduces the risk of death which means more chances of survival. When Hazard Ratio >1 there is less chance of survival probability. Figure 4.6 shows the variables where HR is greater than 1 and less than 1. The bars on the right side represent high effect and vice-versa.

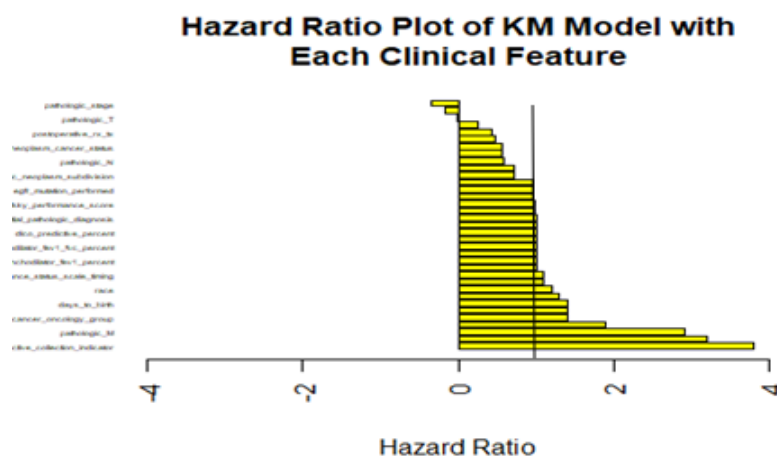


Figure 4.6 KM model hazard ratio plot

Beta: The beta coefficients represent the expected change in the dependent variable for a one-unit change in a predictor variable, holding all other predictors constant. If β coefficient is equal to 0 it means that the variable is strongly correlated. If β coefficient is less than 0 then the variable is not correlated. If the β coefficient is greater than zero the variable is correlated. If β coefficient is a negative value, then the variable is inversely correlated.

The below Figure 4.7 shows the plot of β coefficient which is greater than 0 and less than 0.

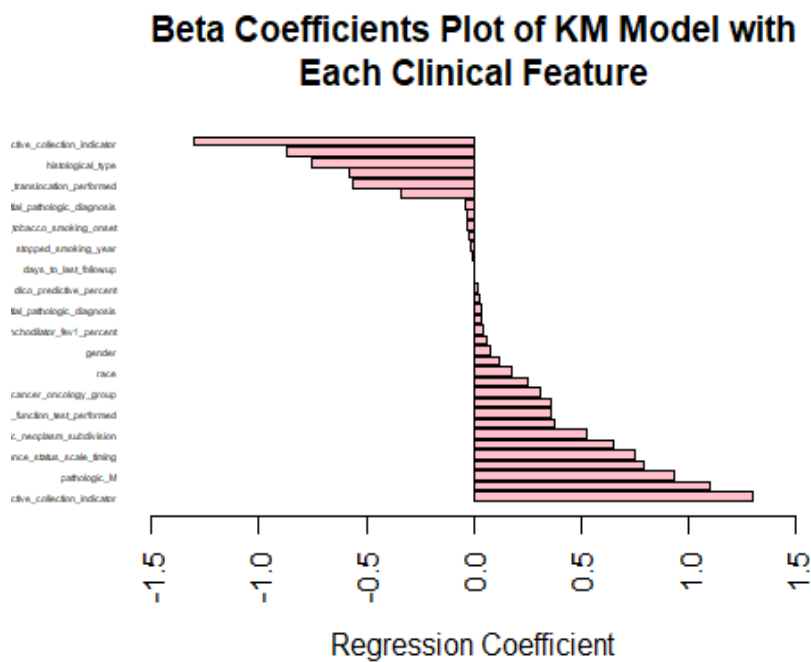


Figure 4.7 KM model beta coefficients plot

Wald Test: the Wald test (also called the Wald Chi-Squared Test, see Table 4.2) is a way to find out about the explanatory variable's significance. If the Wald test shows that the parameters for certain explanatory variables are zero, the variables can be removed from the model. If the test shows the parameters are not zero, the variables should be included in the model. Explanatory variables in a model are significant.

In survival analysis, the hazard ratio (HR) is a measure commonly used to assess the effect of a specific variable or treatment on the risk of an event occurring. It compares the hazard rates between two groups or categories. The hazard rate represents the instantaneous rate at which events (such as death or failure) occur at a given time, given that the individual has survived up to that time. To interpret the hazard ratio, it is essential to consider the HR itself, as well as the lower and upper confidence limits associated with it. Here's an example to illustrate the interpretation: Suppose we are conducting a survival analysis to evaluate the effect of a new drug (Drug A) compared to a standard treatment (Drug B) on the risk of disease progression in cancer patients. The hazard ratio is estimated to be 0.75, with a lower confidence limit of 0.60 and an upper confidence limit of 0.90.

Interpretation:

1. Hazard Ratio (HR): The hazard ratio of 0.75 indicates that patients receiving Drug A have a 25% lower hazard or risk of disease progression compared to those receiving Drug B. In other words, the patients in the Drug A group have a lower likelihood of experiencing the event (disease progression) compared to the Drug B group.
2. Hazard Ratio Lower (HR lower): The lower confidence limit of 0.60 suggests that, with 95% confidence, the true hazard ratio lies between 0.60 and 1.00. This implies that, based on the observed data, patients receiving Drug A have a statistically significant lower hazard of disease progression compared to those receiving Drug B. The lower limit being below 1 indicates a potential benefit of Drug A over Drug B.
3. Hazard Ratio Upper (HR upper): The upper confidence limit of 0.90 indicates that, with 95% confidence, the true hazard ratio lies between 0.90 and 1.00. Although the upper limit includes 1, it is still below 1, suggesting a favourable trend toward reduced risk of disease progression with Drug A compared to Drug B. However, the effect is not statistically significant at the specified confidence level.

In summary, the hazard ratio provides an estimate of the relative risk of an event occurring between two groups. The lower and upper confidence limits indicate the range within which the true hazard ratio is likely to lie, allowing for an assessment of statistical significance and the precision of the estimated effect.

4.3.2 Cox Proportional Hazard (CPH) Model

CPH model is built for all the types of features one by one including clinical, image, and combined (clinical + image). For each type of data, a regression-based process is adopted for the fine-tuning of the CPH model. Stepwise regression implies the building of a regression model following a step-by-step repetitive process. It considers the choosing of independent variables which need to be used in the final model. In this procedure, explanatory variables are inserted or discarded and testing is performed to check their significance at every iteration. There are different methods available to carry out the selection task such as forward selection, backward elimination, and bidirectional elimination [61].

In forward selection, initially, there are no variables, and incrementally, new variables are inserted or deleted. Those having more statistical significance are kept and the process is repeated till reaching optimal results. On the other hand, in backward elimination, at a time, one independent variable is deleted from the entire set, and testing is done to check the significance of deleted variables. Lastly, bi-directional elimination includes both the said methods to check the significance of variables to determine which should be added or discarded [61].

In this work, regression is performed step by step in both directions i.e. bi-directional using two fine-tuning steps. The first is done based on Akaike Information Criteria (AIC), which estimates the prediction error. Considering backward elimination of stepwise regression, it takes the entire model and keeps on deleting variables one by one, and then forward selection is used to select and insert the variables. The procedure results in the final model. The second is done based on the p-value of

each variable. P-value computes the probability of observing the test statistic. The variable with the highest p-value is determined and is deleted. Again, the model is created and this checking is continued until all the variables have a value less than 0.005. At last, the clinical and image features are merged directly.

- **CPH model with clinical data only**

First, we developed the CPH Model with the clinical dataset. The CPH Model was fine-tuned with stepwise regression in both directions (backward and forward) as explained. Table 4.3 and Figure 4.8 show the CPH Model features which were significant. Figure 4.9 shows the output obtained for days_to_death and survival probability using the CPH model with the clinical dataset.

Table 4.3 CPH model with a clinical dataset

Term	estimate	expCOEF	p.value
days_to_last_followup	-0.0023031	0.9976995	0.0000000
karnofsky_performance_score	-0.0270817	0.9732817	0.0000000
post_bronchodilator_fev1_percent	0.0780675	1.0811957	0.0000001
tissue_retrospective_collection_indicatorYES	1.9397013	6.9566727	0.0000004
Race	0.6229740	1.8644647	0.0000029
dlco_predictive_percent	0.0302103	1.0306712	0.0000965
pre_bronchodilator_fev1_percent	-0.0452813	0.9557286	0.0001132
number_pack_years_smoked	-0.0064494	0.9935713	0.0079699
ethnicity	1.0525914	2.8650660	0.0136394

Estimate: Cox regression analysis is a technique for assessing the association between variables and survival rate. The measure of risk provided for each variable is the risk ratio (RR). A risk ratio of 1 means that the risk is the same for each participant. A risk ratio greater than 1 indicates increased risk; a ratio less than 1 indicates less risk.

expCOEF: For a continuous covariate, $\text{Exp}(b)$ is the increase of the hazard ratio for 1 unit change of the continuous variable. When b is negative, then $\text{Exp}(b)$ is less than 1 and $\text{Exp}(b)$ is the decrease

of the hazard ratio for 1 unit change of the continuous variable. For a dichotomous covariate, $\text{Exp}(b)$ is the hazard ratio.

Figure 4.8 shows the variables that are significant. If the P value is greater than 0.005 then it is considered to be significant.

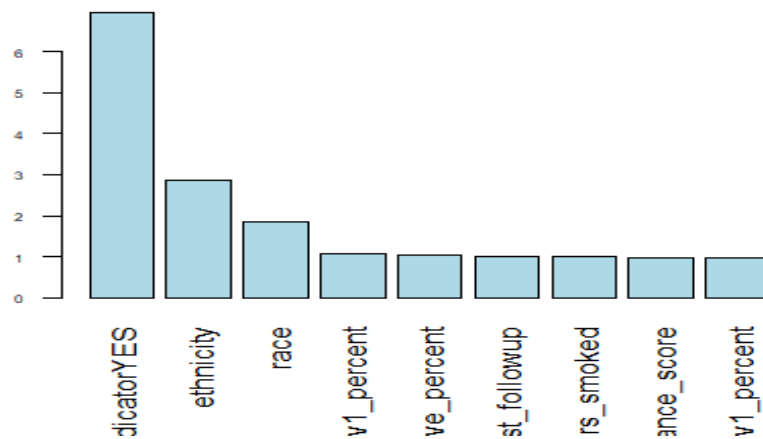


Figure 4.8 Representation of significant variables

Figure 4.9 is the Output obtained for days_to_death and survival probability using the CPH model with the clinical dataset. These lines represent the mean estimates (middle line), the upper bound at a 95% confidence interval (CI), and the lower bound at a 95% confidence interval.

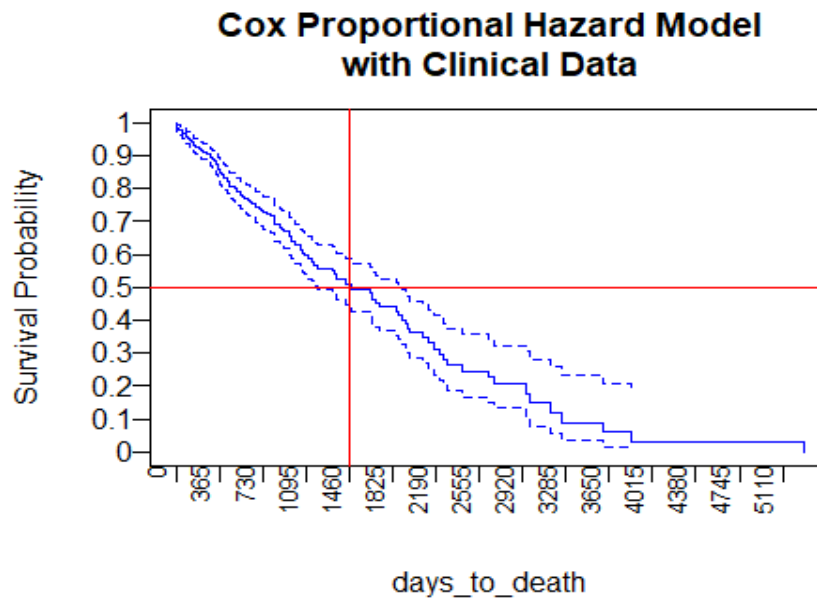


Figure 4.9 Output obtained for days_to_death and survival probability using the CPH model with the clinical dataset.

- **CPH model with image features only**

We developed the CPH Model with an image features dataset. A similar process is followed to fine-tune the CPH Model, as explained. Table 4.4 and Figure 4.10 show the CPH Model features with image data. Figure 4.11 shows the output obtained for days_to_death and survival probability using the CPH model with the clinical dataset.

Table 4.4 CPH model with an image dataset

term	estimate	expCOEF	p.value
q99%	0.0279967	1.028392e+00	0.0000001
sd	-0.2993061	7.413324e-01	0.0000001
gabs_apt_99%	0.0062651	1.006285e+00	0.0000097
gabs_mag_5%	16.0452872	9.297789e+06	0.0000143
gabs_apt_85%	-0.0172230	9.829245e-01	0.0000218
max	0.0132601	1.013348e+00	0.0000927
gabs_apt_15%	-0.2354696	7.901997e-01	0.0000993
gabs_apt_70%	0.0220214	1.022266e+00	0.0001274
gabor_apt_mean	-0.0338855	9.666822e-01	0.0002359
gabor_mag_mean	14.9201309	3.018078e+06	0.0003408
gabs_mag_25%	-10.9231057	1.800000e-05	0.0007260
gabs_apt_55%	0.0244262	1.024727e+00	0.0019221
q85%	-0.0136215	9.864709e-01	0.0031263
q95%	0.0073312	1.007358e+00	0.0031924
gabs_mag_20%	11.3373776	8.389973e+04	0.0054963
gabs_mag_15%	-10.4142989	3.000000e-05	0.0084310
gabs_mag_55%	-4.8069897	8.172400e-03	0.0104240
gabs_mag_70%	-3.6260540	2.662100e-02	0.0113171
gabs_apt_40%	0.0737808	1.076571e+00	0.0142306

Figure 4.10 shows the variables that are significant. If the P value is greater than 0.005 then it is considered to be significant and if the P value is less than 1 then it is less significant.

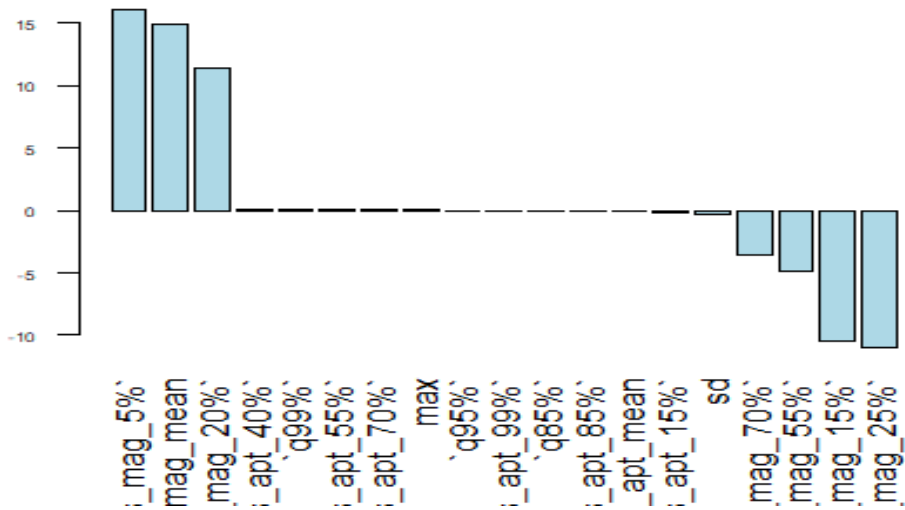


Figure 4.10 Representation of significant variables

Figure 4.11 is the Output obtained for days_to_death and survival probability using the CPH model with the image dataset. These lines represent the mean estimates (middle line), the upper bound at a 95% confidence interval (CI), and the lower bound at a 95% confidence interval.

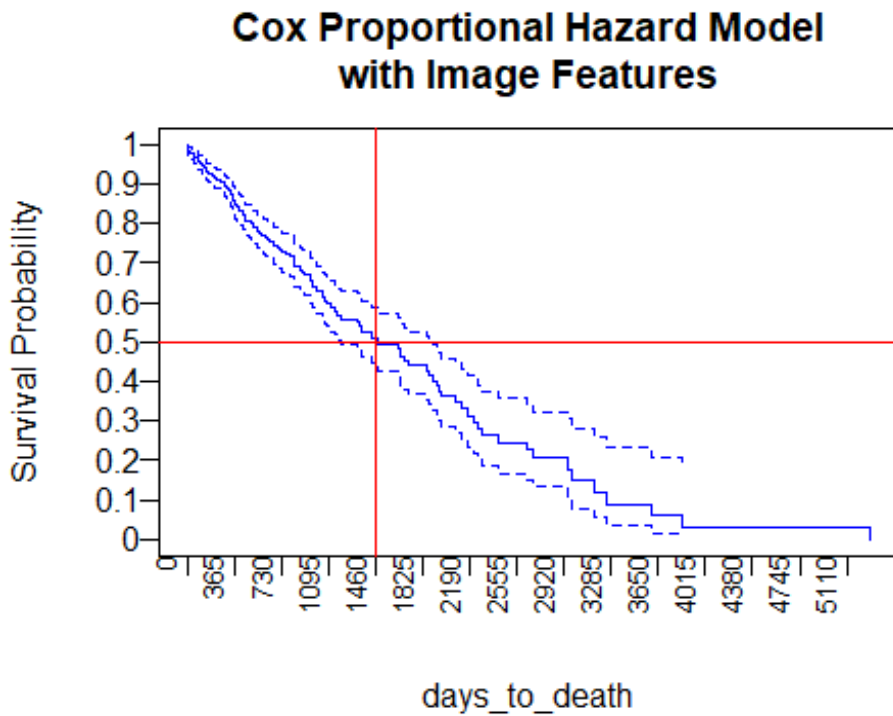


Figure 4.11 Output obtained for days_to_death and survival probability using CPH model with image dataset.

- **CPH model with clinical + image features**

We developed the CPH Model with clinical data and image features dataset. The CPH Model was fine-tuned using similar bidirectional stepwise regression. The following table and figure show the CPH Model features for combined image and clinical data. Table 4.5 and Figure 4.12 show the CPH Model with images and clinical data.

Table 4.5 CPH model with combined clinical and image features

Term	estimate	expCOEF	p.value
days_to_last_followup	-0.0025301	0.9974731	0.0000000
karnofsky_performance_score	-0.0284355	0.9719649	0.0000000
tissue_prospective_collection_indicatorYES	-2.0968787	0.1228393	0.0000001
gabs_mag_70%	-1.0147658	0.3624873	0.0000011
gabs_apr_40%	0.0129594	1.0130438	0.0000011
dlco_predictive_percent	0.0353065	1.0359372	0.0000015
post_bronchodilator_fev1_percent	0.0590699	1.0608494	0.0000418
Race	0.4991652	1.6473455	0.0001834
histological_typeCSCC_Spec	-1.7843745	0.1679020	0.0028685
pre_bronchodilator_fev1_percentj	-0.0305136	0.9699473	0.0115835

Figure 4.12 shows the variables that are significant. If the P value is greater than 0.005 then it is considered to be significant and if the P value is less than 1 then it is less significant.

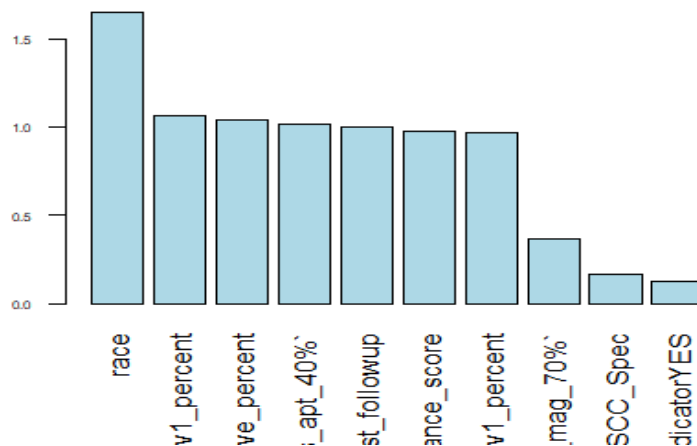


Figure 4.12 Representation of significant variables

Figure 4.13 shows the output obtained for days_to_death and survival probability using the CPH model with clinical and image datasets. These lines represent the mean estimates (middle line), the upper bound at a 95% confidence interval (CI), and the lower bound at a 95% confidence interval.

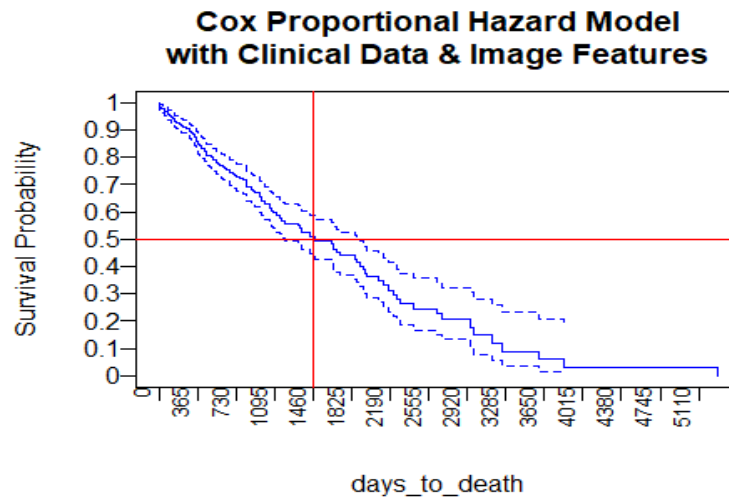


Figure 4.13 Output obtained for days_to_death and survival probability using the CPH model with clinical and image datasets.

From the above model comparisons, it is analyzed that the CPH model performed the same on all three data models yielding identical results. Thus, the results obtained showed the significance of each data in the survival analysis of lung cancer and can be utilized based on different scenarios.

4.3.3 Accelerated Failure Time (AFT) model

To apply AFT models, we must remove zero values from the days_to_death variable.

- **AFT model with clinical data only**

First, we developed the AFT Model with the clinical dataset. Figure 4.14 shows the output obtained for days_to_death and survival probability using the AFT model with clinical data. These lines represent the mean estimates (middle line), the upper bound at a 95% confidence interval (CI), and the lower bound at a 95% confidence interval.

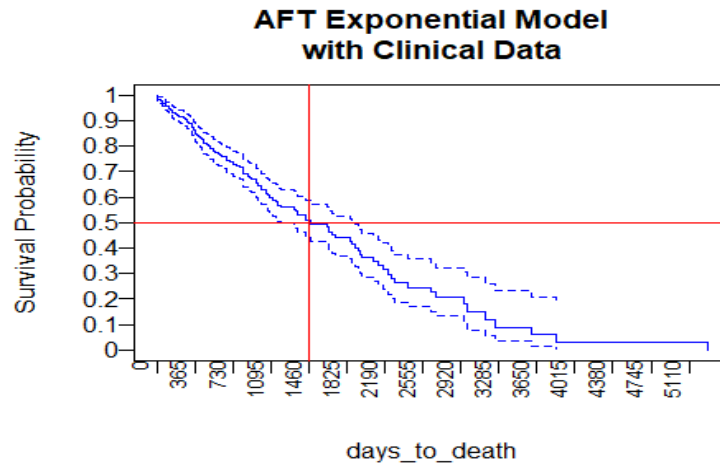


Figure 4.14 Output obtained for `days_to_death` and survival probability using the AFT model with clinical data.

- **AFT model with image data only**

First, we developed the AFT Model with the clinical dataset. Figure 4.15 shows the output obtained for `days_to_death` and survival probability using the AFT model with image data. These lines represent the mean estimates (middle line), the upper bound at a 95% confidence interval (CI), and the lower bound at a 95% confidence interval.

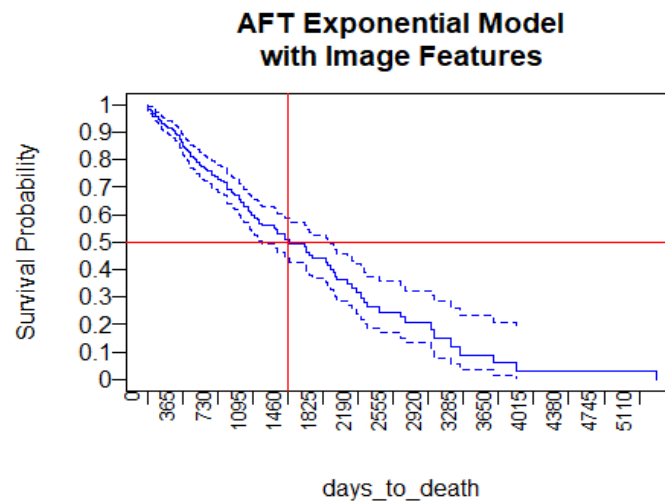


Figure 4.15 Output obtained for `days_to_death` and survival probability using AFT model with image data.

- **AFT model with clinical + image data**

After evaluating the image and clinical data separately, we developed the AFT Model with a

combination of clinical data and image features dataset. Figure 4.16 shows the output obtained for days_to_death and survival probability using the AFT model with clinical + image data. These lines represent the mean estimates (middle line), the upper bound at a 95% confidence interval (CI), and the lower bound at a 95% confidence interval.

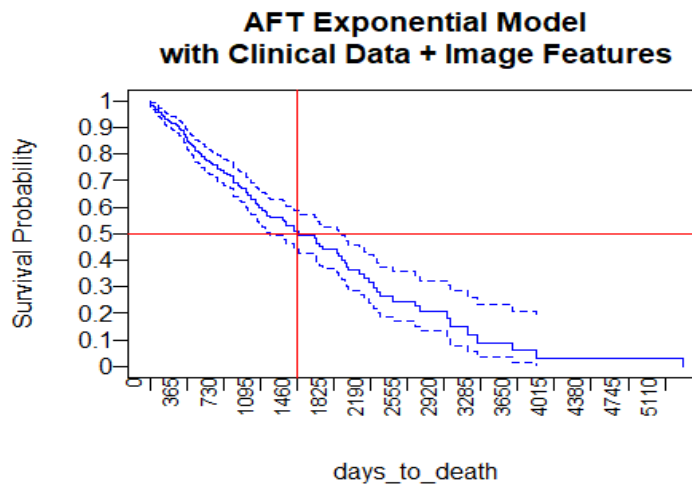


Figure 4.16 Output obtained for days_to_death and survival probability using the AFT model with clinical and image data.

Again, the comparison of AFT model data revealed the significance of clinical, image, and combined data equally. From the analysis of results, it can be inferred that all the considered types of data can be utilized to perform survival analysis of lung cancer.

4.4 Comparison of Models

ANOVA test is recommended to make the comparison between the two models. In this section, the following combination of models was compared and analyzed.

4.4.1 CPH Model Comparison with different data

Three CPH models were prepared using clinical data, image data, and a combined dataset (clinical data + image data).

- Comparing `cph_model` and `cph_model_clinical`

In `cph_model`, both clinical data and image feature data were used and in `cph_model_clinical` only clinical data was utilized.

```

Analysis of Deviance Table
Cox model: response is Surv(days_to_death, vital_status)
Model 1: ~ ethnicity + tissue_retrospective_collection_indicator + karnofsky_performance_score + race + days_to_last_followup + pre_bronchodilator_fev1_percent + post_bronchodilator_fev1_percent + dlco_predictive_percent + number_pack_years_smoked
Model 2: ~ tissue_prospective_collection_indicator + karnofsky_performance_score + race + histological_type + days_to_last_followup + pre_bronchodilator_fev1_percent + post_bronchodilator_fev1_percent + dlco_predictive_percent + `gabs_mag_70%` + `gabs_apt_40%`
loglik  Chisq Df P(>|Chi|)
1 -655.48
2 -646.08 18.786 1 1.462e-05

1
2 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*'
0.05 '.' 0.1 ' ' 1

```

Figure 4.17 Comparison Output

The result obtained in Figure 4.17 shows a degree of freedom (Df) of 1 (indicating that the more complex model has one additional parameter), and a very small p-value ($< .001$). This means that adding the image features data with clinical data to the model did lead to a significantly improved fit over the `cph_model_clinical`.

- **Comparing `cph_model_clinical` and `cph_model_img`**

In `cph_model_clinical`, only clinical data was used and in `cph_model_img` only clinical data was utilized.

```

Analysis of Deviance Table
Cox model: response is Surv(days_to_death, vital_status)
Model 1: ~ ethnicity + tissue_retrospective_collection_indicator + karnofsky_performance_score + race + days_to_last_followup + pre_bronchodilator_fev1_percent + post_bronchodilator_fev1_percent + dlco_predictive_percent + number_pack_years_smoked
Model 2: ~ max + sd + `q85%` + `q95%` + `q99%` + gabor_mag_mean + `gabs_mag_5%` + `gabs_mag_15%` + `gabs_mag_20%` + `gabs_mag_25%` + `gabs_mag_55%` + `gabs_mag_70%` + gabor_apt_mean + `gabs_apt_15%` + `gabs_apt_40%` + `gabs_apt_55%` + `gabs_apt_70%` + `gabs_apt_85%` + `gabs_apt_99%`
loglik  Chisq Df P(>|Chi|)
1 -655.48
2 -781.21 251.47 10 < 2.2e-16

1
2 ***

```

```

---
Signif. codes:
  0 '***' 0.001 '**' 0.01 '*'
  0.05 '.' 0.1 ' ' 1

```

Figure 4.18 Comparison Output

The result obtained in Figure 4.18 shows a Df of 10 (indicating that the more complex model has 10 additional parameters), and a very small p-value ($< .001$). This means that the model trained on image features data did lead to a significantly improved fit over the `cph_model_clinical`.

- **Comparing `cph_model` and `cph_model_img`**

In `cph_model`, we used both clinical data and image features data, and in `cph_model_img` we used only img features data.

```

Analysis of Deviance Table
Cox model: response is Surv(days_to_death, vital_status)
Model 1: ~ max + sd + `q85%` + `q95%` + `q99%` + gabor_mag_mean + `gabs_mag_5%` + `gabs_mag_15%` + `gabs_mag_20%` + `gabs_mag_25%` + `gabs_mag_55%` + `gabs_mag_70%` + gabor_apt_mean + `gabs_apt_15%` + `gabs_apt_40%` + `gabs_apt_55%` + `gabs_apt_70%` + `gabs_apt_85%` + `gabs_apt_99%`
Model 2: ~ tissue_prospective_collection_indicator + karnofsky_performance_score + race + histological_type + days_to_last_followup + pre_bronchodilator_fev1_percent + post_bronchodilator_fev1_percent + dlco_predictive_percent + `gabs_mag_70%` + `gabs_apt_40%`
loglik  Chisq Df P(>|Chi|)
1 -781.21
2 -646.08 270.26 9 < 2.2e-16

1
2 ***
---
Signif. codes:
  0 '***' 0.001 '**' 0.01 '*'
  0.05 '.' 0.1 ' ' 1

```

Figure 4.19 Comparison Output

The result obtained in Figure 4.19 shows a Df of 9 (indicating that the more complex model has 9 additional parameters), and a very small p-value ($< .001$). This means that adding the image features data with clinical data to the model did lead to a significantly improved fit over the `cph_model_img`. From the above tests presented in Figures 4.17-4.19 for clinical (Model 1), image (Model 2) and combined (Model 3), the following analysis and interpretation are made:

- **Comprehensive Information:** Model 3 leverages a broader range of information by incorporating both clinical data and X-ray images. This combination might capture a more comprehensive understanding of the disease, potentially leading to a more robust and generalizable model.
- **Redundancy Reduction:** By combining multiple sources of information, Model 3 may mitigate redundancies present in individual datasets. This reduces the risk of overfitting and improves the model's ability to generalize to new and unseen data.
- **Improved Robustness:** X-ray images and clinical data could complement each other, compensating for limitations or noise present in either dataset alone. The combined model might be more resilient to errors or biases inherent in any single type of data.
- **Enhanced Feature Representation:** X-ray images provide visual information that might capture patterns or features not present or easily quantifiable in clinical data alone. The combined model benefits from a richer feature representation, potentially improving its discriminatory power.
- **Clinical Applicability:** Combining clinical data with imaging might result in a more interpretable model, providing insights that align better with clinical practice. This could enhance the model's utility for healthcare professionals in decision-making scenarios.

These points highlight the potential advantages of Model 3 despite its identical accuracy to Models 1 and 2. While accuracy is an essential metric, considering the broader context, richness of information and practical applicability can often provide a more comprehensive assessment of model performance. In this work, considering the robustness of the data, image features data has been preferred over the clinical and combined to perform further classification of images with cancer.

4.5 Classification Results

This section provides the classification results for three models applied in this work to classify CT scan images as normal or having cancer. It requires sufficient RAM and Processing support to

execute the model smoothly and quickly. We used 4GB RAM and Intel Processor 5th Generation HP Laptop for the task. Firstly, the results for the proposed deep learning model i.e. CNN provided and the other two ensemble learning techniques i.e. RF and XgBoost are evaluated in comparison to CNN. The outcome achieved with all three models provided with the more accurate classifier with less error rate.

4.5.1 Convolutional Neural Network (CNN) Results

The results achieved using CNN reveal a high validation accuracy of 99% in classifying the images of valid data sets. The trendlines in Figure 4.20 show how the model builds epoch by epoch in terms of accuracy and loss. The less the loss, the better the model is and the more the accuracy is, the better the model is considered.

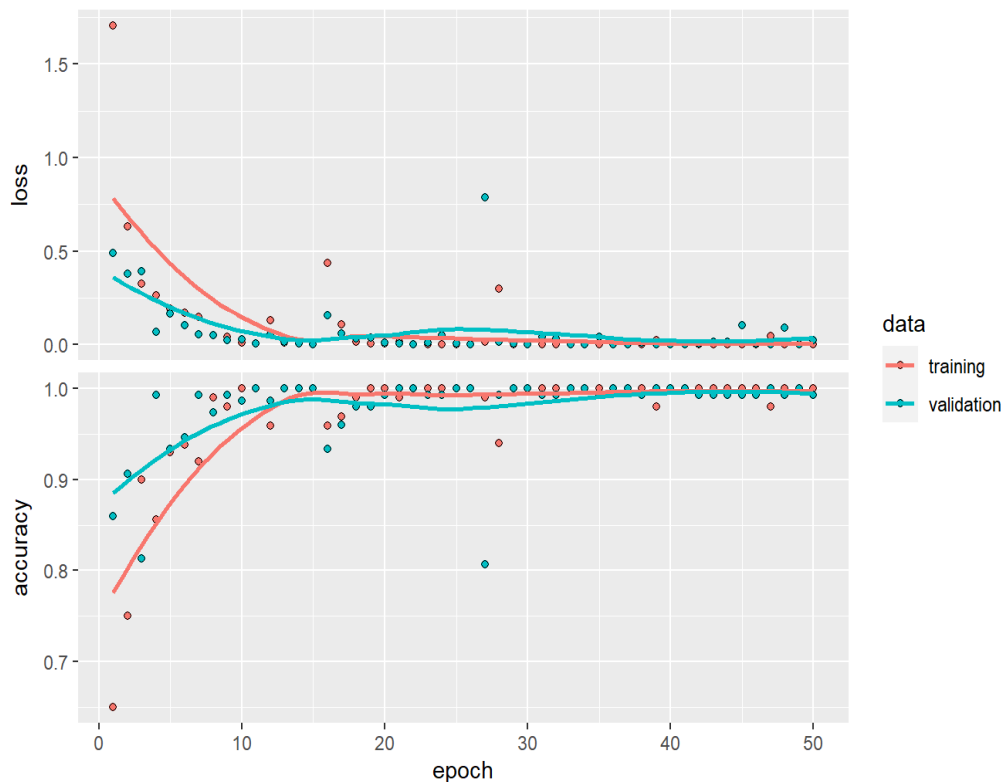


Figure 4.20 Accuracy and Loss Results for CNN

The blue line represents the CNN model performance on training data and the red line represents the CNN model performance on validation data. If these two lines run in parallel, the CNN model is assumed to be of excellent standard. This is achieved by the CNN model. If trendlines diverge, the model is considered unstable. The following section in Figure 4.21 shows the accuracy of the CNN

model is 99% which means out of 100 images, 99 images are correctly classified by the CNN model and one image is miss-classified.

```
## Final epoch (plot to see history):  
## loss: 0.0003888  
## accuracy: 1  
## val_loss: 0.02403  
## val_accuracy: 0.9933
```

Figure 4.21 Results for CNN

4.5.2 Random Forest (RF) Results

The following section in Figure 4.22 shows the results statistics achieved by applying the RF model. The screenshot shows that the implication of RF on the considered data yielded an accuracy rate of 95.83%.

```
Confusion Matrix and Statistics  
  
## Reference  
## Prediction  0  1  
##           0 10  0  
##           1  3 59  
## Accuracy : 0.9583  
## 95% CI : (0.883, 0.9913)  
## No Information Rate : 0.8194  
## P-Value [Acc > NIR] : 0.0004623  
## Kappa : 0.8453  
## McNemar's Test P-Value : 0.2482131  
##           Sensitivity : 1.0000  
##           Specificity : 0.7692  
##           Pos Pred Value : 0.9516  
##           Neg Pred Value : 1.0000  
##           Prevalence : 0.8194  
##           Detection Rate : 0.8194  
##           Detection Prevalence : 0.8611
```

```
##          Balanced Accuracy : 0.8846
##          'Positive' Class : 1
```

Figure 4.22 Results for RF

4.5.3 Xgboost Results

The following section in Figure 4.23 shows the results statistics achieved by applying the RF model.

The screenshot shows that the implication of XgBoost on the considered data yielded an accuracy the same accuracy rate as RF i.e. 95.83%.

```
## Confusion Matrix and Statistics
##          Reference
## Prediction  0  1
##          0 10  0
##          1  3 59
##          Accuracy : 0.9583
##          95% CI : (0.883, 0.9913)
##          No Information Rate : 0.8194
##          P-Value [Acc > NIR] : 0.0004623
##          Kappa : 0.8453
##          McNemar's Test P-Value : 0.2482131
##          Sensitivity : 0.7692
##          Specificity : 1.0000
##          Pos Pred Value : 1.0000
##          Neg Pred Value : 0.9516
##          Prevalence : 0.1806
##          Detection Rate : 0.1389
##          Detection Prevalence : 0.1389
##          Balanced Accuracy : 0.8846
##          'Positive' Class : 0
```

Figure 4.23 Results for Xgboost

Therefore, the overall results achieved demonstrate the highest performance and accuracy rate with the proposed CNN in comparison to RF and XgBoost models. These results can be effectively utilized by clinicians to perform better lung diagnosis.

4.6 Comparison with Previous Studies

To validate the success and efficacy, it is important to evaluate the performance of the proposed model against previous research works. Therefore, following the purpose, the presented multi-layered CNN model is compared with the state-of-the-art studies that utilized different techniques for the diagnosis of lung cancer. Due to higher adoptability, accuracy is utilized as the prime evaluation measure to compare proposed and other techniques. A brief overview of such a comparison is presented in Table 4.6 and can be graphically visualized in Figure 4.24. Chen et al.

Table 4.6 Accuracy comparison of the proposed model with state-of-the-art methods

Reference	Year	Dataset	Method/Technique	Accuracy
Shayesteh, S.P. [32]	2020	CT images (n=59)	Logistic based method	61.0%
Huang, B [33]	2022	PET, CT data (n=965)	CNN, RSF	79.0%
Bebas, E. [37]	2021	PET, MR lung images (n=44)	SVM, KNN, RF, and DL	75.4% (SVM)
Chen and Dhahbi[56]	2021	TCGA	Lasso, Xgboost, Proposed biomarker method	92.0%
Dehkharghanian et al [57]	2023	TCGA	2 Deep Neural Networks (DNN) i.e. DenseNet121, KimiaNet	86.0% (KimiaNet)
Sun et al [58]	2019	TCGA	DNN model	97.4%
Ramirez et al [59]	2020	TCGA	Graph CNN	94.7%
Qiu et al [60]	2017	TCGA	--	95.0%
Proposed Study		TCIA	Proposed Multi-layered CNN, RF, and XGBoost	99%

Chen and Dhahbi, 2021 [56] proposed a study to perform classification using an overlapping-based feature selection algorithm. The experimentation is conducted on the TCGA dataset using proposed and other models including Lasso, Xgboost, etc. The results obtained revealed the highest accuracy of 92% with the proposed biomarkers approach. A research work by Dehkharghanian et al. [57] employed two Deep Neural Networks (DNN) namely DenseNet121 and KimiaNet to perform feature extraction and classification for lung cancer detection. The findings obtained highlighted the efficiency of KimiaNet with an accuracy of more than 86% in comparison to DenseNet121 with an accuracy rate of 70%. Similarly, in another study by Sun et al. [58], a DNN model is applied i.e. Genome deep learning to investigate the relationships among variations in genome and traits. Twelve types of cancers are distinguished from the 1991 healthy tissues based on the proposed model and achieved the best accuracy of 97.4%. A technique of Graph CNN is proposed in a study by Ramirez et al. [59] to perform accurate detection of lung cancer at early stages. The TCGA dataset is utilized to differentiate the cancer into 33 types and normal considering 10,340 and 731 cancerous and normal tissue samples, respectively. The employed model achieved an excellent accuracy rate of 94.7% for accurate cancer prediction. Similarly, evaluating the TCGA database for image-based lung cancer analysis and noting the genome pattern dissimilarities, Qiu et al. [60] attained an accuracy of 95%.

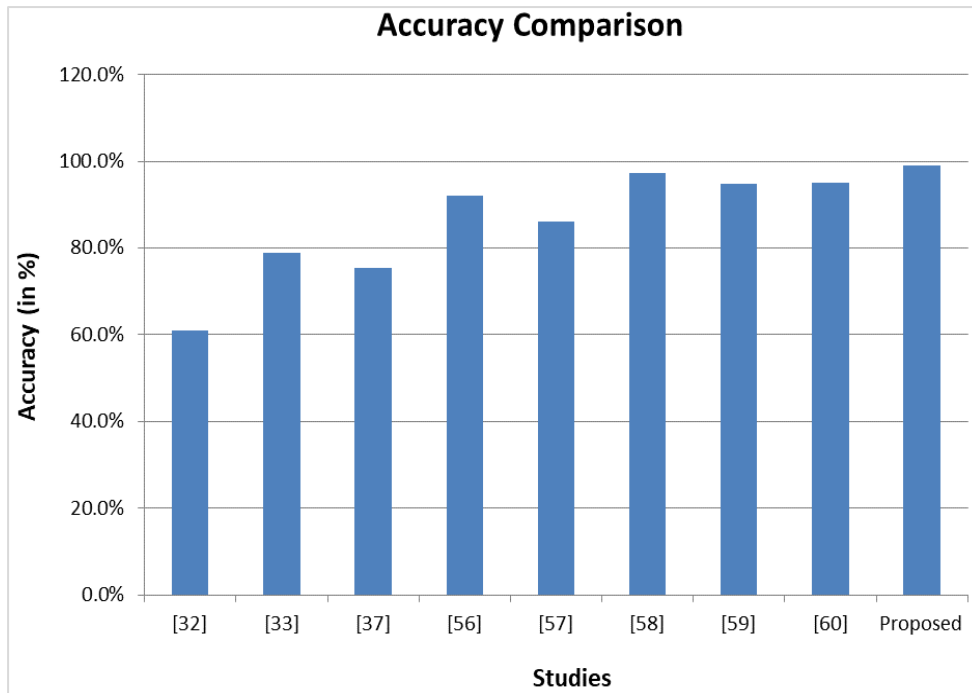


Figure 4.24 Accuracy comparison of the Proposed and Previous studies

Apart from the TCGA dataset, some researchers used CT, and PET imaging for the classification of lung cancer. A study by Shayesteh et al. [32] used radiomic data considering CT images of 59 persons to detect lung cancer. The implication of a logistic-based model achieved an accuracy of 61% in such analysis. Another study by Huang et al. [33] using CNN and RSF methods based on PET and CT data attained a satisfactory accuracy of 79%. Different models including KNN, SVM, RF, and DL are evaluated in a research work by Besbes et al. [37] on PET and MR lung image data. The highest accuracy of 75.4% is demonstrated by SVM in lung cancer classification.

To enhance the diagnostic accuracy of lung cancer using advanced approaches such as deep learning, the present work is proposed. The lack of evaluation of clinical, image, and combined data for lung cancer analysis using statistical and learning-based models motivated the interest of this research to perform such analysis. The results presented in Table 4.5 indicate that, with an outstanding accuracy of 99%, the proposed CNN technique has outperformed the state-of-the-art. Therefore, the proposed strategy has a strong potential to perform robust analysis of lung cancer at early stages and can help clinicians suggest better treatment

Chapter 5

Conclusions and Future Work

Investigating the survival of cancer patients is critical for disease control and treatment method evaluation. It is necessary to find and analyze the risk factors involved in lung cancer to enable its early diagnosis. Therefore, this thesis work proposed to perform an in-depth analysis of crucial variables involved in the growth of lung cancer using statistical models and applied an efficient deep learning model to classify images with cancer.

5.1 Conclusions

We applied three statistical models namely Kaplan Meier (KM), Cox Proportional Hazards (CPH), and Accelerated Failure Time (AFT) were applied to check the significance level of variables on the survival rate of cancer. The analysis was performed considering different types of data i.e. clinical only, image only, and clinical and image combination.

The implication of three statistical models revealed the effect of different factors on the survival of lung cancer based on the p-value which can be focused on performing more efficient lung cancer analysis. Further, the evaluation of results obtained from statistical analysis demonstrated the significance of all the types of data i.e. clinical, image, and combined. The combination of clinical and image data might capture a more comprehensive understanding of the disease, potentially leading to a more robust and generalizable model and may mitigate redundancies present in individual datasets. This reduces the risk of overfitting and improves the model's ability to generalize to new and unseen data. Despite the benefits, the combined model might be more resilient to errors or biases inherent in any single type of the acquired data. X-ray images provide visual information that might capture patterns or features not present or easily quantifiable in clinical data alone. The combined model benefits from a richer feature representation,

potentially improving its discriminatory power. Also, combining clinical data with imaging might result in a more interpretable model, providing insights that align better with clinical practice. This could enhance the model's utility for healthcare professionals in decision-making scenarios.

For classification purposes, an efficient multilayer deep learning model was adopted namely Convolutional Neural Network (CNN) having less computational cost and time overhead. It consists of different layers starting from taking input to producing the output. Finally, the results of CNN were compared with other models. The results obtained using CNN were compared with the two other ensemble approaches i.e. Random Forest (RF) and XgBoost based on accuracy. An accuracy of 99% was achieved using CNN in classifying the images which was higher than the accuracy rates of Random Forest (RF) and XgBoost i.e. 95.83% and 95.83% as well as from the previous studies. This work performed statistical analysis as well as ML-based evaluation of data for the survival analysis of lung cancer and to predict the sample images with cancer. Further, the efficacy of different types of data is evaluated using Kaplan Meier, Cox Proportional Hazard, and Accelerated Failure Time models to determine their significance in the considered analysis. Finally, the CNN prediction model obtained from this case study proves to be robust and stable. The designed CNN model can be, therefore, used on new CT scan images for lung cancer diagnosis for further research and can be helpful for clinicians.

5.2 Future Work and Recommendations

This thesis work contributed to the performance of survival analysis of lung cancer patients based on clinical and image features using statistical models and an effective Deep Learning Technique. However, based on the analysis, there are some points listed below that can be explored and provide a way to conduct further research.

- To perform analysis by applying more complex deep learning approaches and hybrid models for better accuracy.

- To consider other cancer datasets and also to check the efficacy of the applied approach in the classification of other types of cancer with lung cancer.
- To consider the patient's data with severity levels to explore the disease in-depth for its early diagnosis.
- To perform optimization of different parameters of the model to yield accurate results within less time.

References

1. National Cancer Institute: Understanding Cancer. <https://www.cancer.gov/about-cancer/understanding/what-is-cancer>
2. World Health Organization: <https://www.who.int/>
3. Canadian Cancer Society: Cancer Statistics at a glance. <https://cancer.ca/en/research/cancer-statistics/canadian-cancer-statistics>
4. Brenner, D.R., Poirier, A., Woods, R.R., Ellison, L.F., Billette, J.M., Demers, A.A., Zhang, S.X., Yao, C., Finley, C., Fitzgerald, N. and Saint-Jacques, N., 2022. Projected estimates of cancer in Canada in 2022. *CMAJ*, 194(17), pp.E601-E607.
5. Cleveland Clinic: Lung Cancer. <https://my.clevelandclinic.org/health/diseases/4375-lung-cancer>
6. Cancer Research Uk: Stages and Grades of Lung Cancer. <https://www.cancerresearchuk.org/about-cancer/lung-cancer/stages-types-grades/stages-grades>
7. Clark, T.G., Bradburn, M.J., Love, S.B. and Altman, D.G., 2003. Survival analysis part I: basic concepts and first analyses. *British journal of cancer*, 89(2), pp.232-238.
8. Raqab, M.Z., Al-Awadhi, S.A. and Kundu, D., 2018. Discriminating among Weibull, log-normal, and log-logistic distributions. *Communications in Statistics-Simulation and Computation*, 47(5), pp.1397-1419.
9. Hastie, D.I., Liverani, S., Azizi, L., Richardson, S. and Stücker, I., 2013. A semi-parametric approach to estimate risk functions associated with multi-dimensional exposure profiles: application to smoking and lung cancer. *BMC medical research methodology*, 13(1), pp.1-13.
10. A Short Course on Survival Analysis. https://bookdown.org/sestelo/sa_financial/the-semiparametric-model.html
11. Goel, M.K., Khanna, P. and Kishore, J., 2010. Understanding survival analysis: Kaplan-Meier estimate. *International journal of Ayurveda research*, 1(4), p.274.

12. Artificial Intelligence, Machine Learning, and Deep Learning. What's the Real Difference?
<https://medium.com/swlh/artificial-intelligence-machine-learning-and-deep-learning-whats-the-real-difference-94fe7e528097>
13. Artificial Intelligence (AI) vs Machine Learning (ML). <https://azure.microsoft.com/en-in/solutions/ai/artificial-intelligence-vs-machine-learning/#introduction>
14. Wakefield, K., 2021. A guide to the types of machine learning algorithms and their applications. https://www.sas.com/en_gb/insights/articles/analytics/machine-learning-algorithms.html
15. Mathew, A., Amudha, P. and Sivakumari, S., 2020, February. Deep learning techniques: an overview. In *International conference on advanced machine learning technologies and applications* (pp. 599-608). Springer, Singapore.
16. Ngo, G., Beard, R. and Chandra, R., 2022. Evolutionary bagging for ensemble learning. *Neurocomputing*, 510, pp.1-14.
17. Wang, H., Xing, F., Su, H., Stromberg, A. and Yang, L., 2014. Novel image markers for non-small cell lung cancer classification and survival prediction. *BMC bioinformatics*, 15(1), pp.1-12.
18. Palani, D. and Venkatalakshmi, K., 2019. An IoT based predictive modelling for predicting lung cancer using fuzzy cluster based segmentation and classification. *Journal of medical systems*, 43(2), pp.1-12.
19. Bhatia, S., Sinha, Y. and Goel, L., 2019. Lung cancer detection: a deep learning approach. In *Soft Computing for Problem Solving* (pp. 699-705). Springer, Singapore.
20. Lakshmanaprabu, S.K., Mohanty, S.N., Shankar, K., Arunkumar, N. and Ramirez, G., 2019. Optimal deep learning model for classification of lung cancer on CT images. *Future Generation Computer Systems*, 92, pp.374-382.

21. Joon, P., Bajaj, S.B. and Jatain, A., 2019. Segmentation and detection of lung cancer using image processing and clustering techniques. In *Progress in advanced computing and intelligent engineering* (pp. 13-23). Springer, Singapore.
22. Talukdar, J. and Sarma, D.P., 2018. A Survey on Lung Cancer Detection in CT scans Images Using Image Processing Techniques. *International Journal of Current Trends in Science and Technology*, 8(3), pp.20181-20186.
23. Kurkure, M. and Thakare, A., 2016, August. Lung cancer detection using genetic approach. In *2016 International Conference on Computing Communication Control and automation (ICCUBEA)* (pp. 1-5). IEEE.
24. Abedi, S., Janbabaei, G., Afshari, M., Moosazadeh, M., Alashti, M.R., Hedayatizadeh-Omran, A., Alizadeh-Navaei, R. and Abedini, E., 2019. Estimating the survival of patients with lung cancer: What is the best statistical model?. *Journal of Preventive Medicine and Public Health*, 52(2), p.140.
25. Alomaish, H., Ung, Y., Wang, S., Tyrrell, P.N., Zahra, S.A. and Oikonomou, A., 2021. Survival analysis in lung cancer patients with interstitial lung disease. *Plos one*, 16(9), p.e0255375.
26. Cui, L., Li, H., Hui, W., Chen, S., Yang, L., Kang, Y., Bo, Q. and Feng, J., 2020. A deep learning-based framework for lung cancer survival analysis with biomarker interpretation. *BMC bioinformatics*, 21(1), pp.1-14.
27. Al Mamlook, R.E., Bzizi, H.F. and Chen, S., 2020, July. Evaluate Performance Risk Score in Patients Suffering from Lung Cancer Using Survival Analysis of Statistics. In *2020 IEEE International Conference on Electro Information Technology (EIT)* (pp. 145-150). IEEE.
28. Nageswaran, S., Arunkumar, G., Bisht, A.K., Mewada, S., Kumar, J.N.V.R., Jawarneh, M. and Asenso, E., 2022. Lung Cancer Classification and Prediction Using Machine Learning and Image Processing. *BioMed Research International*, 2022.
29. Wang, J., Chen, N., Guo, J., Xu, X., Liu, L. and Yi, Z., 2021. SurvNet: A novel deep neural network for lung cancer survival analysis with missing values. *Frontiers in Oncology*, 10, p.588990.

30. Yu, K.H., Zhang, C., Berry, G.J., Altman, R.B., Ré, C., Rubin, D.L. and Snyder, M., 2016. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nature communications*, 7(1), pp.1-10.
31. Wang, S., Liu, Z., Chen, X., Zhu, Y., Zhou, H., Tang, Z., Wei, W., Dong, D., Wang, M. and Tian, J., 2018, July. Unsupervised deep learning features for lung cancer overall survival analysis. In *2018 40th Annual international conference of the IEEE engineering in medicine and biology society (EMBC)* (pp. 2583-2586). IEEE.
32. Shayesteh, S.P., Shiri, I., Karami, A.H., Hashemian, R., Kooranifar, S., Ghaznavi, H. and Shakeri-Zadeh, A., 2020. Predicting lung cancer Patients' survival time via logistic regression-based models in a quantitative radiomic framework. *Journal of Biomedical Physics & Engineering*, 10(4), p.479.
33. Huang, B., Sollee, J., Luo, Y.H., Reddy, A., Zhong, Z., Wu, J., Mammarrappallil, J., Healey, T., Cheng, G., Azzoli, C. and Korogodsky, D., 2022. Prediction of lung malignancy progression and survival with machine learning based on pre-treatment FDG-PET/CT. *EBioMedicine*, 82, p.104127.
34. Zeng, Y., Cao, W., Wu, C., Wang, M., Xie, Y., Chen, W., Hu, X., Zhou, Y., Jing, X. and Cai, X., 2022. Survival Prediction in Home Hospice Care Patients with Lung Cancer Based on LASSO Algorithm. *Cancer Control*, 29, p.10732748221124519.
35. She, Y., Jin, Z., Wu, J., Deng, J., Zhang, L., Su, H., Jiang, G., Liu, H., Xie, D., Cao, N. and Ren, Y., 2020. Development and validation of a deep learning model for non-small cell lung cancer survival. *JAMA network open*, 3(6), pp.e205842-e205842.
36. Deist, T.M., Dankers, F.J., Ojha, P., Marshall, M.S., Janssen, T., Faivre-Finn, C., Masciocchi, C., Valentini, V., Wang, J., Chen, J. and Zhang, Z., 2020. Distributed learning on 20 000+ lung cancer patients—The Personal Health Train. *Radiotherapy and Oncology*, 144, pp.189-200.
37. Bębas, E., Borowska, M., Derlatka, M., Oczeretko, E., Hładuński, M., Szumowski, P. and Mojsak, M., 2021. Machine-learning-based classification of the histological subtype of non-small-cell lung cancer using MRI texture analysis. *Biomedical Signal Processing and Control*, 66, p.102446.

38. Gu, Q., Feng, Z., Liang, Q., Li, M., Deng, J., Ma, M., Wang, W., Liu, J., Liu, P. and Rong, P., 2019. Machine learning-based radiomics strategy for prediction of cell proliferation in non-small cell lung cancer. *European journal of radiology*, 118, pp.32-37.
39. The Cancer Imaging Archive (TCIA) Public: The Cancer Genome Atlas Lung Squamous Cell Carcinoma Collection (TCGA-LUSC) Access.<https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=16056484>.
40. Platias, C. and Petasis, G., 2020, September. A comparison of machine learning methods for data imputation. In *11th Hellenic Conference on Artificial Intelligence* (pp. 150-159).
41. Stekhoven, D.J. and Bühlmann, P., 2012. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1), pp.112-118.
42. Dudley, W.N., Wickham, R. and Coombs, N., 2016. An introduction to survival statistics: Kaplan-Meier analysis. *Journal of the advanced practitioner in oncology*, 7(1), p.91.
43. Lin, D.Y. and Wei, L.J., 1989. The robust inference for the Cox proportional hazards model. *Journal of the American statistical Association*, 84(408), pp.1074-1078.
44. Saikia, R. and Barman, M.P., 2017. A review on accelerated failure time models. *Int J Stat Syst*, 12(2), pp.311-322.
45. Efron, B., 1988. Logistic regression, survival analysis, and the Kaplan-Meier curve. *Journal of the American statistical Association*, 83(402), pp.414-425.
46. Fisher, L.D. and Lin, D.Y., 1999. Time-dependent covariates in the Cox proportional-hazards regression model. *Annual review of public health*, 20(1), pp.145-157.
47. Zeng, D. and Lin, D.Y., 2007. Efficient estimation for the accelerated failure time model. *Journal of the American Statistical Association*, 102(480), pp.1387-1396.
48. O'Shea, K. and Nash, R., 2015. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*.

49. Albawi, S., Mohammed, T.A. and Al-Zawi, S., 2017, August. Understanding of a convolutional neural network. In *2017 international conference on engineering and technology (ICET)* (pp. 1-6). Ieee.
50. Kaggle: CT Medical Images. <https://www.kaggle.com/datasets/kmader/siim-medical-images>.
51. Sakib, S., Ahmed, N., Kabir, A.J. and Ahmed, H., 2019. An overview of convolutional neural network: its architecture and applications.
52. Dai, B., Chen, R.C., Zhu, S.Z. and Zhang, W.W., 2018, December. Using random forest algorithm for breast cancer diagnosis. In *2018 International Symposium on Computer, Consumer and Control (IS3C)* (pp. 449-452). IEEE.
53. Strobl, C., Malley, J. and Tutz, G., 2009. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological methods*, *14*(4), p.323.
54. Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H. and Chen, K., 2015. Xgboost: extreme gradient boosting. *R package version 0.4-2*, *1*(4), pp.1-4.
55. Elkan, C., 2012. Evaluating classifiers. *San Diego: University of California*.
56. Chen, J.W. and Dhabhi, J., 2021. Lung adenocarcinoma and lung squamous cell carcinoma cancer classification, biomarker identification, and gene expression analysis using overlapping feature selection methods. *Scientific reports*, *11*(1), p.13323.
57. Dehkharghanian, T., Bidgoli, A.A., Riasatian, A., Mazaheri, P., Campbell, C.J., Pantanowitz, L., Tizhoosh, H.R. and Rahnamayan, S., 2023. Biased data, biased AI: deep networks predict the acquisition site of TCGA images. *Diagnostic pathology*, *18*(1), pp.1-12.
58. Sun, Y., Zhu, S., Ma, K., Liu, W., Yue, Y., Hu, G., Lu, H. and Chen, W., 2019. Identification of 12 cancer types through genome deep learning. *Scientific reports*, *9*(1), p.17256.

59. Ramirez, R., Chiu, Y.C., Hererra, A., Mostavi, M., Ramirez, J., Chen, Y., Huang, Y. and Jin, Y.F., 2020. Classification of cancer types using graph convolutional neural networks. *Frontiers in physics*, 8, p.203.
60. Qiu, Z.W., Bi, J.H., Gazdar, A.F. and Song, K., 2017. Genome-wide copy number variation pattern analysis and a classification signature for non-small cell lung cancer. *Genes, Chromosomes and Cancer*, 56(7), pp.559-569.
61. Smith, G., 2018. Step away from stepwise. *Journal of Big Data*, 5(1), pp.1-12.

Appendix A

- **Code snippets of Random Forest (RF)**

Some of the code snippets of RF are given in the screenshot:

```
df_train_normal$status=c(rep(0,132))  
df_train_cancer$status=c(rep(1,465))  
df_val_normal$status=c(rep(0,13))  
df_val_cancer$status=c(rep(1,59))
```

```
df_train=rbind.data.frame(df_train_normal, df_train_cancer)  
df_val=rbind.data.frame(df_val_normal, df_val_cancer)
```

```
sam=sample(1:nrow(df_train), replace = FALSE, 1*nrow(df_train))  
df_train=df_train[sam,]  
sam_val=sample(1:nrow(df_val), replace = FALSE, 1*nrow(df_val))  
df_val=df_val[sam_val,]
```

```
library(randomForest)
```

```
## randomForest 4.7-1.1
```

```
model_rf=randomForest(x = df_train[, colnames(df_train) != "status"],  
                      y = as.factor(df_train$status))
```

```
model_rf
```

```
##  
## Call:  
## randomForest(x = df_train[, colnames(df_train) != "status"], y = as.factor(df_train$status))  
##           Type of random forest: classification  
##           Number of trees: 500  
## No. of variables tried at each split: 150  
....
```

```

...
##           OOB estimate of  error rate: 0.5%
## Confusion matrix:
##      0   1 class.error
## 0 132   0 0.000000000
## 1   3 462 0.006451613

```

```

preds_rf=predict(model_rf, newdata = df_val, type="prob")

preds_rf=as.data.frame(preds_rf)

preds_fact=as.factor(ifelse(preds_rf$`1`>0.5,1,0))

library(caret)

```

```

con_rf=confusionMatrix(preds_fact, as.factor(df_val$status), positive = "1")

con_rf

```

- **Code snippets of XgBoost**

Some of the code snippets of XgBoost are given in the screenshot:

```

train_features=df_train[,1:22500]
train_labels=as.data.frame(df_train[,22501])
y=as.numeric(train_labels$`df_train[, 22501]`)
val_features=df_val[,1:22500]
val_labels=as.data.frame(df_val[,22501])
val_y=as.numeric(val_labels$`df_val[, 22501]`)

library(xgboost)

dtrain <- xgb.DMatrix(data =as.matrix(train_features), label= y)
set.seed(1111)

model.xgb <- xgboost(data = dtrain, # the data
                     max.depth = 3,
                     nround = 50, # boosting iterations
                     objective = "binary:logistic")

## [1] train-logloss:0.450405
## [2] train-logloss:0.312112
## [3] train-logloss:0.221966
## [4] train-logloss:0.161211
## [5] train-logloss:0.118934

```

```
## [6] train-logloss:0.090433
## [7] train-logloss:0.068632
## [8] train-logloss:0.052441
## [9] train-logloss:0.040700
## [10] train-logloss:0.032173
## [11] train-logloss:0.025885
## [12] train-logloss:0.021088
## [13] train-logloss:0.017334
## [14] train-logloss:0.014195
## [15] train-logloss:0.011987
## [16] train-logloss:0.010363
## [17] train-logloss:0.009064
## [18] train-logloss:0.007998
## [19] train-logloss:0.007154
## [20] train-logloss:0.006434
## [21] train-logloss:0.005957
## [22] train-logloss:0.005579
## [23] train-logloss:0.005250
## [24] train-logloss:0.004956
## [25] train-logloss:0.004695
## [26] train-logloss:0.004469
## [27] train-logloss:0.004271
## [28] train-logloss:0.004084
## [29] train-logloss:0.003931
## [30] train-logloss:0.003789
## [31] train-logloss:0.003644
## [32] train-logloss:0.003525
## [33] train-logloss:0.003410
## [34] train-logloss:0.003410
## [35] train-logloss:0.003410
## [36] train-logloss:0.003410
## [37] train-logloss:0.003410
## [38] train-logloss:0.003410
## [39] train-logloss:0.003410
## [40] train-logloss:0.003410
## [41] train-logloss:0.003410
## [42] train-logloss:0.003410
```

```
## [43] train-logloss:0.003410
## [44] train-logloss:0.003410
## [45] train-logloss:0.003410
## [46] train-logloss:0.003410
## [47] train-logloss:0.003410
## [48] train-logloss:0.003410
## [49] train-logloss:0.003410
## [50] train-logloss:0.003410

preds_xgb=predict(model.xgb, newdata=xgb.DMatrix(data =as.matrix(val_features))
, type="prob")

preds_xgb=as.data.frame(preds_xgb)

colnames(preds_xgb)="Prob"

pred_xgb_levels_1=as.factor(ifelse(preds_xgb$Prob>0.5,1,0))

confusionMatrix(pred_xgb_levels_1, as.factor(val_y))
```