COMPARING PREDICTION ACCURACIES OF CANCER SURVIVAL USING MACHINE LEARNING TECHNIQUES AND STATISTICAL METHODS IN COMBINATION WITH DATA REDUCTION METHODS

A Dissertation
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Mohammad Gulam Mostofa

In Partial Fulfillment of the Requirements
for the Degree of
DOCTOR OF PHILOSOPHY

Major Department:
Statistics

June 2022

Fargo, North Dakota

North Dakota State University
Graduate School

Title

Comparing Prediction Accuracies of Cancer Survival Using Machine Learning Techniques and Statistical Methods in Combination with Data Reduction Methods

By

Mohammad Gulam Mostofa

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

DOCTOR OF PHILOSOPHY

SUPERVISORY COMMITTEE:

Dr. Megan Orr
Chair

Dr. Rhonda Magel

Dr. Changhui Yan

Dr. Mingao Yuan

Approved:

July 8, 2022
Date

Dr. Rhonda Magel
Department Chair

ABSTRACT

This comparative study of five-year survival prediction for breast, lung, colon, and leukemia cancers using a large SEER dataset along with 10-fold cross-validation provided us with an insight into the relative prediction ability of different machine learning and data reduction methods. Lasso regression and the Boruta algorithm were used for variables selection, and Principal Component Analysis (PCA) was used for dimensionality reduction. We used one statistical method Logistic regression (LR) and several machine learning methods including Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), K Nearest Neighbor (KNN), Artificial Neural Network (ANN), and Naïve Bayes Classifier (NB). For breast cancer, we found LDA, RF, and LR were the best models for five-year survival prediction based on the accuracy, sensitivity, specificity, and area under the curve (AUC) using data reduction method from Z score normalization and the Boruta algorithm. The results for lung cancer indicated the SVM linear, RF, and ANN were the best survival prediction models using data reduction methods from the Z score and max min normalization. The results for colon cancer indicated, ANN, and RF were the best prediction models using the Boruta algorithm and Z score method. The results for leukemia showed ANN, and the RF were the best survival prediction models using the Boruta algorithm and data reduction technique from the Z score. Overall, ANN, RF, and LR were the best prediction models for all cancers using variables selection by the Boruta algorithm.

ACKNOWLEDGEMENTS

I am grateful to Almighty Allah for the good health and well-being that were necessary to complete this research. I would like to thank the Department of Statistics at North Dakota State University for all the support I was given to excel in my studies. I would like to thank Professor Dr. Mingao Yuan for enormous help in the methodology sections to continue my research. I wish to express my profound gratitude to my Ph.D. advisor Dr. Orr for her advice, patience, and suggestions to complete my research. She helped me a lot with the writing of this paper. My sincere appreciation goes to Professor Dr. Magel, for her continuous suggestions, financial support, and for making resources available during my study. I would like to thank my committee member, Dr. Changhui Yan, for his encouragement, insight, and comments.

I wish to thank my wife Dr. Fatima, my son Mohiuddin, my daughter Mubassira, my parents, my mother-in-law, and my father-in-law for their love, prayer, and support throughout my studies. It would have been impossible to finish my research and dissertation without their support.

DEDICATION

I dedicate this dissertation to my wife, my kids, and my parents.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I: INTRODUCTION

Cancer is abnormal cell growth with the potential to spread or invade other parts of the body, damaging normal cells *(Anand et al., 2008)*. The two main types of cancer are solid tumor cancers and hematologic cancers. Hematologic cancers are related to blood cells and include leukemia, lymphoma, and multiple myeloma (www.cancer.org). Solid tumor cancers form a lump called a tumor, which is malignant *(Anand et al., 2008).* Another type of tumor is a benign tumor, but these tumors never spread out in the body and are not associated with cancer. Common solid tumor cancers are related to any part of a body organ or tissue and include breast, prostate, lung, and colorectal cancers. Signs and symptoms of solid tumor cancers include, among others, abnormal bleeding, a lump, inexplicable weight loss, and continuous cough. Tobacco is the leading cause of cancer death in the United States (www.cancer.gov/about-cancer/causes-prevention/risk/tobacco). Other causes of cancer death include a poor diet, obesity, excessive alcohol drinking, lack of physical activity, and genetic predisposition. According to the Surveillance, Epidemiology, and End Results Program (SEER) under the National Cancer Institute, the estimated number of new cases of cancer will be 1.9 million and the estimated number of cancer deaths will be 609,360 in 2022 (Siegel et.al, 2022). Cancer is a heterogeneous disease including several diverse subtypes.

In cancer research, the early diagnosis and prognosis of a cancer type are very important for the clinical management of patients. Epidemiological measures can be estimated for all cancers based on an individual's age, gender, race, sex, socio-economic status, and other factors (e.g. smoking habit, family history, health conditions) (Bartholomai & Frieboes, 2018). In this study, we will focus on the prediction accuracy of five-year survival for four types of cancer: breast, lung,

colon, and leukemia. The motivation for our research was the lack of literature on comprehensive studies for comparing five-year survival prediction accuracy among machine learning techniques**.**

1.1. Lung Cancer

Lung cancer is the leading cause of cancer-related death in the United States and is common in both males and females. There are four kinds of lung cancer: large cell carcinomas, small cell carcinomas, squamous cells, and adenocarcinomas. Smoking is the cause of most lung cancers. Thus, due to smoking, the relative risk of getting lung cancer in smokers is higher than in non-smokers (Doll et al., 2004). In the United States, lung cancer incidence also varies due to racial disparities. Lung cancer control and proper decision-making regarding treatment are always challenging for patients, doctors, and other personnel. The intent of computing survival accuracy is of strong importance in providing information and improving care to patients and clinicians. From the dataset of lung cancer patients with demographic (e.g., age), diagnostic (e.g., tumor size), and procedural information (e.g., Radiation and/or Surgery applied), the question is whether patient survival accuracy can be computationally predicted with any precision.

1.2. Breast Cancer

Breast cancer is the second leading cause of death among women, but it is quite rare in men and is more common in middle-aged women than young women (Gupta, et.al. 2011). According to SEER, the estimated number of new female breast cancer cases will be 287,850 and an estimated number of 43,250 people will die of this disease in 2022. Based on 2016-2018 data, about 12.9% of women will be diagnosed with breast cancer during their lifetime. Risk factors for breast cancer include inheritance, age, and lifestyle behaviors such as exercise and diet. Breast cancer is identified either by mammogram screening or by perceiving a lump in the breast (Sharma et al., 2018). Breast cancer has been studied worldwide to improve survival by focusing on

reducing risks, finding causes, developing new diagnostic, and developing new treatment protocols (www.cancer.org/cancer/breast-cancer). The diagnosis of breast cancer at an earlier stage is very significant because treatment and early diagnosis help to prevent the spreading of breast cancer.

1.3. Colon Cancer

Colon cancer ranks third in the United States according to the number of cancer diagnoses in both women and men. In 2022, an estimated 106,180 cases of colon cancer will be diagnosed in the US, and a total of 52,580 people will die from these cancers. The colon cancer incidence rate dropped by 1% from 2013-to 2017 due to changing their lifestyle-related risk factors and more people were getting screened (https://www.cancer.org/cancer/colon-rectal-cancer/about/key-statistics.html), Surgery is the most common treatment for colon cancer that has not spread to distant sites. Colon cancer patients typically receive chemotherapy after surgery. Therefore, prediction of survival and screening can prevent colon cancer through the detection and removal of precancerous growths (polyps), as well as to detect of cancer at an early stage, when treatment is usually more successful and less intensive.

1.4. Leukemia Cancer

Leukemia is a cancer of the bone marrow and blood-forming tissue including the lymphatic system and usually involves the white blood cells. The body produces large numbers of abnormal blood cells if leukemia has developed. The abnormal cells are white blood cells in most cases of leukemia, The leukemia cells do not function properly since they are different from normal blood cells. In the United States, leukemia is diagnosed in about 2000 children and 29,000 adults (https://training.seer.cancer.gov/leukemia/intro). In 2022, an estimated 60,650 new cases of leukemia will be diagnosed in the US and 24,000 people will die from the disease ((Siegel et al., 2020). Chemotherapy, sometimes in combination with targeted drugs, is used to treat most acute

leukemias. Treatment advances such as the development of targeted drugs have resulted in large survival improvements for most types of leukemia.

1.5. Predicting Cancer Outcomes with Machine Learning Techniques

The prediction of disease outcomes is a challenging and interesting task for physicians. Physicians have access to massive amounts of data needed to compare treatment outcomes for all cancer types, but they still need to analyze that information and blend it with a patient's medical profile. Therefore, physicians can get more information about cancers fiver year survival prediction results based on the accuracy, sensitivity, and specificity. (Vickers, 2011). The emergence of large cancer datasets available and collected for the medical community and researchers due to the advent of new technologies have led to an expansion of techniques used to analyze medical data, including Machine Learning (ML) techniques. Machine Learning is a type of Artificial Intelligence (AI) that helps to make predictions, estimations, and decisions based on a large dataset (www.cancer.gov/research/areas/diagnosis/artificial-intelligence).

The main goal for ML is the discovery of new facts from large datasets based on logical and statistical methods. ML is used in models associated with cancer survival, prognosis, reliability estimates, and better accuracy (Liou & Chang, 2015). There are two main types of ML algorithms, supervised learning, and unsupervised learning, that are used to build a mathematical relationship between the inputs and desired outputs of a data set. There are a variety of supervised learning techniques, including Decision Tree (DT), Random Forest (FR), Logistic Regression (LR), Support Vector Machine (SVM), Artificial Neural Network (ANN), KNN (K-Nearest Neighbor), and Naïve Bayes (NB). Unsupervised learning methods include Principal Component Analysis (PCA), Hierarchical Clustering, K-means Clustering, and Independent Component Analysis.

A machine learning technique generally involves two steps. First, a subset of the sample data, called a training set, is used as an input to build a model. Second, after building the training model, the remaining data, called the testing set, are used for testing the utility of the training model. The model is expected to predict the output using the testing set (Salod & Singh, 2019).

For cancer survival prediction, previously published papers have used SEER datasets (from 1973 to 2001), while other papers have used different cancer datasets (e.g. University of California Irvine online database (Hong & Yang, 1991). Wisconsin Breast Cancer dataset (Street et al., 1993)and The Digital Database for Screening Mammography (Heath et al., 1998). To our knowledge, no study has evaluated the SEER dataset from 2004 to 2016 for cancer survivability prediction using ML techniques. Moreover, there is no comprehensive study using statistical modeling and ML techniques for the prediction of cancer survival prediction. A few studies used modern ML techniques for cancer survival prediction, but they did not consider all models together to compare prediction accuracy. In this study, we use newer as well as more established ML techniques to predict five-year cancer survival. We also investigate the effect of multiple data reduction techniques on these predictions.

1.6. Research Objectives

There are three main objectives for this study:

1. To predict five-year cancer survivability using machine learning (ML) and statistical techniques.
2. Investigating the prediction accuracy of the techniques using different data reduction techniques
3. To identify the best cancer survival prediction models based on the accuracy, sensitivity, specificity, and area under the curve**.**

1.7. Organization

The rest of this dissertation is organized as follows. We will discuss a literature review in chapter II, methodology in chapter III, the results for breast cancer in chapter IV, the results for lung cancer in chapter V, the results for colon cancer in chapter VI, the results for leukemia cancer in chapter VII, and conclusion and research contribution in chapter VII.

CHAPTER II: LITERATURE REVIEW

2.1. Methods Used to Analyze SEER Data Set for Cancer Prediction

Delen et al. (2005) conducted a study about predicting breast cancer survivability by comparing three data mining techniques using SEER breast cancer data from 1973 to 2000. First and foremost, the authors requested data files through the SEER website (www.seer.cancer.gov). The SEER cancer data consists of nine text files that are related to cancer for a specific anatomical site such as breast, colon, female genital, lymphoma, male genital, respiratory, urinary, and leukemia. There are 72 variables in each file and each file relates to a specific incidence of cancer. These 72 variables furnish cancer-specific and socio-demographic information. The authors used SPSS statistical analysis tool and statistical data miner to manipulate the data. They compared the breast cancer five-year prediction accuracies of decision trees, artificial neural networks, and logistic regression. For the neural network analysis, only one hidden layer was considered. Accuracy computed by an ANN was 91.2%. The other machine learning technique, decision tree, outperformed the other methods with 93.6 % accuracy. The authors used some mathematical algorithms such as information gain, Gini index, and entropy to construct a tree to improve the prediction accuracy as well as the C5 algorithm as their decision tree method. One statistical technique used in this study was logistic regression, which assumes that the response variable is binary and thus predicts the odds of its occurrence. Accuracy computed by logistic regression was 89.2%.

Bellaachia and Guven (2006) carried out a study about predicting breast cancer survivability using data mining techniques. The authors carried out three data mining techniques for breast cancer survival prediction: LR, ANN, and the C4.5 decision tree algorithms. The authors compared these three techniques based on survival prediction accuracy. According to their

findings, the C4.5 method was much better than that of the logistic regression and Neural network. The study demonstrated that three machine learning techniques are very promising for survivability prediction. The prediction accuracy of ANN, C4.5, and logistic regression were 91.21%, 93.26%, and 89.20% respectively.

Mourad et al. (2020) investigated feature selection and machine learning techniques for thyroid cancer prognosis. The researcher used the SEER thyroid cancer data from 1988 to 2007. The researchers used several predictor variables such as age, gender, race, tumor size, grade, stages, primary disease extent, location of nodal disease, and a few positive lymph nodes. This study demonstrated an artificial neural network for predicting thyroid cancer survival with non-parametric methods for feature selection such as Fisher's discriminant ratio and Kruskal-Wallis. Ten years survival prediction accuracy of the neural network was 94.5 %.

Jajroudi et al. (2014) reported prediction of survival in thyroid cancer using data mining techniques. The researchers used the SEER thyroid cancer data from 1973 to 2000. The variables included tumor size, grade, pathologic stage, lymph node sub-type record, and RX sum-Surg. Logistic Regression, decision tree, and ANN models were used to predict survival in thyroid cancer and compare accuracy. According to the results, the decision tree represents a good model of five-year survival predictions in thyroid cancer patients with 93.6 % accuracy. The prediction accuracy of ANN was 91.2% and the accuracy of logistic regression was 89.2 %.

Fradkin et al. (2006) used Support Vector Machine (SVM) and penalized logistic regression to construct the predictive model for lung cancer survival, and analyzed the important features based on model parameters. The authors used nine variables including age, sex, race, place of birth, histology, diseases of extent, radiation, surgery, and causes of death. The authors used the SEER lung cancer dataset from 1973 to 2002. Their experiment suggests that SVM provided

better results than penalized logistic regression. Several variables were considered based on their ranking of prediction contribution. The rank of prediction contribution was selected based on sensitivity and specificity.

Rajesh and Anand (2012) implemented the C4.5 classification algorithm for breast cancer prediction. The researchers used the SEER data from 1973 to 1998. Six variables were used for this study such as CS extension, Age, Regional nodes positive, sequence number, and CS tumor size. This study was considered ten years survival prediction. The authors used KNN, C4.5, and NB machine learning techniques. The study showed 94% accuracy for the C4.5 model, 93% accuracy for KNN, and 92% accuracy for the NB classifier.

Agrawal et al. (2012) reported lung cancer prediction using the SEER data from 1998 to 2008. The researchers used the supervised classification methods decision tree, random forest, and ensemble model to predict the survival of lung cancer patients at the end of 6 months, 9 months, 1 year, 2 years, and 5 years of survival. The decision tree performed best based on the area under the ROC curve and accuracy for lung cancer survival. Using the ensemble voting classification scheme, prediction accuracies of 73.61%, 74.45%, 76.80%, 85.45%, and 91.35% were obtained for the 6-month, 9-month, 1-year, 2-year, and 5-year lung cancer survival prediction. SEER datasets from 1977 to 1988 for the analysis of survival. He developed a novel encoding of good and poor prognosis of censored data in an ANN architecture to provide a framework for prognostic prediction. This paper shows a new approach to prognostic prediction using a neural network. The ANN was applied to breast cancer prognosis resulting in accurate models which play a role in preventing unnecessary surgeries.

Burke et al. (1997) compared the TNM staging system’s predictive accuracy with that of ANN for the 5-year survival of patients. The National Cancer Institute's SEER breast carcinoma

data was collected from 1977-to 1982. The researchers made the comparison over three different datasets such as SEER data, PCE data, and PCE colorectal dataset. The ANN predictions of the 5-year survival of patients with breast carcinoma were significantly more accurate than the TNM staging system (ANN, 0.770; TNM, 0.720; $P < 0.001$). The artificial neural network's predictions of 10-year survival were significantly more accurate than the TNM staging system (ANN, 0.730; TNM, 0.692; $P < 0.01$). ANN was more accurate than the TNM staging system in both cases.

2.2. Methods Used to Analyze Other Data Sets for Cancer Survival Prediction

Ganggayah et al., (2019) investigated predictive factors for the survival of breast cancer patients using machine learning techniques. The authors collected a breast cancer data set consisting of 8,492 patient records from the University Malaya Medical Centre (UMMC), Kuala Lumpur, Malaysia with diagnosis information between 1993 and 2016 for this study. Initially, 112predictor variables and one response variable were in this data set, but after discussions with several clinicians in UMMC about predictor variables, 89 variables that were unnecessary for breast cancer survival were removed from the dataset. The final data set contained 8,066 patient records with 23 predictor variables and one response variable. The response variable in this study was the survival status of the patients. The prediction models were built through random forest, decision tree, neural network, logistic regression, adaptive boosting, and support vector machine techniques. The important variables were selected via the random forest. The highest prediction accuracy was 82.7 % for random forest and the lowest was obtained from the decision tree (accuracy= 79.8%).

Ming et al.(2019) investigated breast cancer risk prediction using machine learning techniques and breast cancer risk assessment tools. The authors used baseline data from a

prospective randomized trial in Michigan. They also used Swiss clinic-based retrospective breast cancer data from the oncology department at the Geneva University Hospital. Several Predictor variables were used including age, race, gender, number of biopsies, atypical, hyperplasia, and deceased status. In this study, data from o112,587 individuals were analyzed using generalized linear models (GLM), logistic regression, linear discriminant analysis (LDA), Quadratic discriminant analysis (QDA), and KNN, and. The predictive accuracy was 88.89 % for random forest and 88.28 % using Adaptive Boosting (AB).

Salod and Singh (2019) explained the performance of machine learning algorithms in breast cancer detection and screening. The authors used breast cancer data from the University of California Irvine (UCI) online database. This dataset contains 116 individuals and ten quantitative predictor variables. Of the 116 individuals, 64 individuals (55 %) belong to the breast cancer tumor present, and 52 individuals (45 %) belong to the breast cancer tumor absent group. The authors considered several machine learning and statistical techniques including logistic regression (LR), SVM, KNN, Decision tree, and boosting algorithms. After each model was trained, it was tested via validation and test sets. Some common metrics such as accuracy, sensitivity, specificity, and Receiver Operating Characteristics (ROC) were used to evaluate the models' performances.

Gupta et.al (2011) summarized various review and technical papers on breast cancer prognosis and diagnosis problems and concluded that for many applied data mining classification techniques, the accuracy of diagnosis was highly acceptable and could help medical professionals in decision-making for early diagnosis and to avoid biopsy. Accuracy was higher for ANN compared to other classification techniques for the prognostic problem.

Lundin et al. (1999) have applied ANN and LR machine learning techniques to compare 5-year, 10-year, and 15-year breast cancer survival prediction using the total number of individuals

951 from the City Hospital of Turku and Turku University Central Hospital. Eight variables were used to include tumor size, axillary nodal status, histological type, mitotic count, nuclear pleomorphism, tubule formation, tumor necrosis, and age. The authors compared logistic regression with ANN based on the AUC. The AUC of logistic regression for 5, 10, and 15 years were 0.897, 0.862, and 0.858 respectively. On the other hand, the AUC of ANN for 5, 10, and 15 years was 0.909, 0.886, and 0.883 respectively. Based on the AUC, ANN yields better results compared to logistic regression.

Abdelaal et al. (2010) investigated the capability of the classification SVM with RF and DF in analyzing the DDSM (The Digital Database for Screening Mammography) dataset for the extraction of the mammographic mass variables along with age that discriminates true and false cases. Several machine learning techniques were used such as SVM, DT, and RF The authors used AUC for identifying the best model. Based on the AUC, SVM (0.79768) was better than that of DT, (0.5388), and RF(0.57575).

Liou et.al (2015) explained prediction models of breast cancer using ANN, DF, LR, and genetic algorithms. The authors collected. data from the University of Wisconsin Breast Cancers. The accuracy for ANN was 0.9878 (Sensitivity= 1, Specificity=0.9802), LR was 0.9434 (Sensitivity= 0.9716, specificity=0.9482), DF was 0.9434 (sensitivity=0.9615, specificity=0.9105) and the genetic algorithm was 0.9502. (sensitivity=0.9602, specificity=0.9273). Based on the results the ANN performed the best.

Burke et al. (1997) compared the 5-year prediction accuracy for breast cancer using ANN with TNM staging system. This study used the Patient Care Evaluation dataset from 1983 to 1992 by the Commission on Cancer of the American College of Surgeons. (www.facs.org/quality-programs/cancer-programs/national-cancer-database/). The dataset contained 54 input variables.

The accuracy for ANN was 0.770 and the TNM staging system was 0.720. Therefore, ANN performed the best model.

Gultepe, Y. (2021) investigated the performance of classification algorithms to predict lung cancer survival. The lung cancer dataset was collected from the Machine Learning Repository website of the University of California, Irvine (Hong & Yang, 1991). The dataset provided many variables including age, gender, alcohol use, genetic risk, chronic lung disease, balanced diet, obesity, smoking, passive smoking, chest pain, blood cough, weight loss, shortness of breath, difficulty swallowing, and snoring, and others. The authors used several machine learning techniques including KNN, DT, SVM, NB, LR, and RF. PCA was used to reduce dimensionality, and the Z score and max-min were used for normalization. The best results were accuracy found from raw data with NB (0.57), KNN (0.71), and DT (0.71), while the worst results were obtained from RF (0.43) algorithm. Based on the Z score, the best results were an accuracy found with KNN (0.83), while the worst results were obtained with DT (0.33). On the other hand, the best accuracy results were obtained from LR and SVM (0.71), while the worst result was found from NB (0.29).

Al-Bahrani et al., (2013) explained one year, two years, and five years of colon cancer survival prediction using SEER data from 1973 to 2009. The authors used various predictor variables including EOD, AJCC stage 3rd ed, birthplace, lymph node involvement, regional node-positive, surg prim site, histology type, behavior, reasons for no surgery, age at diagnosis, tumor size, and primary site. Several ML techniques such as RF, DT, LR, and Ensemble Voting were used in this study. The ensemble method had reached the highest accuracy of 90.38% for one year, 88.01% for 2 years, and 85.13% for five-year survival. Furthermore, the AUC for one year was 0.96, for two years for 0.95, and 0.92 for five years for the ensemble voting method.

Hassouneh et al. (2019) explained the leukemia survival prediction using SEER data from 1973 to 2014. The researchers used three machine learning techniques including DT, ANN, SVM, and one Deep Neural Network (DNN). The accuracy of DNN was about 75%, DT and SVM were about 73.45%, and ANN was about 74%. Therefore, DNN was the best model rather than the three ML techniques.

Table 1. Summary of datasets and machine learning methods from related article

	Data set analyzed		Methods used to predict cancer survival										
Articles	SEER	Other	DT	ANN	LR	Naïve Bayes	C4.5	SVM	RF	GBM	AB	KNN	LDR
Delen et al. (2005)	X		X	X	X								
Bellaachia and Guven (2005)	X		X	X		X	X						
Chip et.al. (2017)	X		X					X	X	X			
Moustafa et.al. (2020)	X			X									
M Jajroudi et.al. (2014)	X			X	X								
Dmitriy Fradkin et.al. (2005)	X							X					
K.Rajesh and Sheila Anand (2012	X					X	X					X	
Ankit Agrawal et.al. (2012)	X		X						X				
W. Nick Street (1998)	X			X									
Harry B. Burke et al (1997)	X			X									
Mogana et.al. (2019)		UMMC	X	X	X			X					
Ching Ming et.al. (2019)		GUH			X				X		X	X	X
Zakia and Yashik (2019)		UCI	X		X			X				X	
Joseph and David (2007)		Review paper	X	X		X		X					
Nikita and Subhalaxmi Das (2020)		TCGA	X	X				X				X	X
M. Lundin et al (1999)		CHTT		X									
Medhat and Muhamed (2010)		DDSM	X					X	X				
Der-Ming and Wei-pin (2015)		WBC	X	X									
G.Thippa Reddy et.al. (2020)		CTG	X			X		X	X				
Burke et al.(1995)		PCE		X									
Yasemin Gültepe (2021	X		X		X				X				X
Al-Bahrani et al., (2013)	X		X	X				X					X

CHAPTER III: METHODS

3.1. Data Description and Source

Data from the Surveillance, Epidemiology, and End Results (SEER) Program was obtained by request through the SEER Program website (www.seer.cancer.gov). The SEER Program collects cancer survival and incidence data from nine participating registries in the United States and makes these datasets available to laboratories and institutions for analytical research. The SEER cancer data consists of text files corresponding to specific anatomical sites such as breast, lung (respiratory), colon and rectum, and leukemia. Each file contains 124 variables, and each record relates to a specific incidence of cancer. The cancer mortality rates and incidence trends in SEER are presumed to be representative of the cancer mortality rates and incidence in the United States (Hankey et al., 1999). These datasets are considered reliable and contain comprehensive information on cancer survival and incidence in the United States (https://www.cancer.gov/about-cancer/managing-care/using-trusted-resources). The National Institute of Health (NIH) organized and gave support to collecting cancer incidence from the population-based cancer registries covering about 34.6 percent of the U.S. population. SEER data from 1973 to 2016 were collected, but we used lung cancer, breast cancer, colon cancer, and leukemia incidence data from 2004 to 2016 for our study(www.seer.cancer.gov).

3.2. Data Processing and Variables Description

The cancer datasets consisted of single flat files with a fixed-width text format and SEER description documentation explaining each variable with unique values. The raw data was uploaded onto SAS and imported into R for processing and analysis. The SEER cancer data consisted of 684,571 cases/records with 124 variables. Pre-processing of the data was performed by removing cases with unknown or missing values. Different numbers indicate missing values

such as 99, 999, 9999, and blank space I have considered all for data cleaning. Since the goal of many machine learning techniques is to develop models for predicting the survival incidence of cancer, the survival variable used herein was encoded as a binary dependent variable with values 0 (survived) and 1 (did not survive). The total number of lung cancer patients was 242,876 in the lung cancer dataset from 2004 to 2016. The clean datasets of 37,844 lung cancer patients' records. The clean datasets of breast cancer, colon cancer, and leukemia cancer consisted of 78,320, 48,447, and 96,227 cases, respectively. The datasets contained 205 predictor variables and one dependent variable (categorical variable) which reflected the survival status of patients. We used 15 main predictor variables in our study because all variables were not useful for cancer incidence such as laterality, gender, year of birth, lymphomas, etc. In addition, some variables in the cancer dataset that contained redundant information such as variable overrides and variable recodes, and these variables were removed from the data set. For example, Morphology and Extent of Disease provide aggregated information on various attributes of cancer. Furthermore, the Morphology variables furnish Behavior, Histology, and Grade Code, each of which consisted of unique information about cancer tumors. Moreover, the Extent of Disease variable provides six different characteristics of the tumor. We chose to use their derivative variables with more detailed information in place of using aggregated variables (Delen et.al. 2005). Therefore, we chose 15 predictor variables that were more consistent for our study. Table 2 lists all variables included in our study.

Table 2. Predictor variables for survival modeling

Categorial Variables Names	SAS Names in SEER data	Number of Unique Values
Marital Status	MAR_STAT	6
Race	RACEIV	28
Stage	DAJCCSTG	5
Surgery	NO_SURG	6
Radiation	RADIATNR	10
Grade	GRADE	5
Behavior	BEHO3V	2
Histology	HISTO3V	91
Extent of Diseases (EOD)	CSEXTEN	29
Lymnodes Involvement	CSLYMPHN	10
Primary site	PRIMSITE	9
Continuous variables		
Age	AGE_DX	110
Number of positive nodes	EOD10_PN	95
Number of regional lymnode	REG_NUM	41
Tumor size	CSTUMSIZ	98

3.2.1. Marital Status

The patient's marital status was determined at the time of diagnosis for the reportable tumor. It is one of the independent prognostic factors for inflammatory breast cancer (Yan-ling et.al. 2019). Unmarried patients were classified as single, unmarried, widowed, divorced, and separated.

3.2.2. Race

The race is mainly divided into black, white, and other (including American Indian/Alaska Native and Asian/Pacific Islander). The race is not a biologically defined parameter but racial differences in cancer outcomes and characteristics have been published (Gadgeel & Kalemkerian, 2003). Higher rates of cancer incidence in black patients compared to white patients have been observed. The highest rate of lung cancer (8.5% risk of lung cancer diagnosis) and lung cancer

mortality rate (7.6% risk of death) occurs in African Americans in the United States (Schoenfeld D and Fraumeni J, 2006).

3.2.3. Stages

The survival stage is a factor describing the extent of cancer spread in the body. The American Joint Committee on Cancer (AJCC) first classified cancer stages by T (the primary tumor), N (regional lymph nodes), and M (distant metastasis). Subsequently, AJCC uses a numeric system to describe cancer stages which started from Stage0 to stage IV. Stage zero indicates in situ which describes cancers are still located in the place they started. They have not spread to nearby tissues. Stage I describes evidence of cancer growth and tissue of origin. Stage II describes cancer signifies a limited local spread. Stage III indicates the extension of local and regional spread. Stage IV describes distant metastasis. In our study, we used cancer stages from Stage I to Stage IV (www.seer.cancer.gov).

3.2.4. Grade

The grade is another important independent variable for survival prediction. The grade code denotes the amount of differentiation from well-differentiated Grade I to undifferentiated Grade IV. The grade is a morphological variable that provides unique information about the tumor.

3.2.5. Additional Morphological Variables

There are other morphological variables, consisting of Histology, Behavior, and Extent of Disease (EOD), that furnish unique information about the tumor. The morphology records the kind of tumor that has developed and how it behaves. The behavior of a tumor is the way it acts within the body. A tumor behavior can be benign or malignant. The extent of the disease coding scheme records the number of regional nodes found positive for cancer at pathological examination. Histology describes the tissue for the primary tumor and the microscopic

composition of cells. It is a basis for the determination of treatment options and staging. The *International Classification of Diseases for Oncology,* Third Edition (ICD-O-3) is referred for coding the histology primary site and identifies the site in which the primary tumor originated. The primary site was coded by the International Classification of Diseases for Oncology, Third Edition (ICD-O-3). The number of primaries counts all tumors that were reportable in the year they were diagnosed even if the tumors occurred before the registry existed, or before the registry participated in the SEER Program. Tumor size is strongly related to prognosis (chances for survival). In general, the smaller the tumor, the better the prognosis tends to be. The tumor's exact size is from 001 to 988. Radiation reflects whether a treatment was external beam, brachytherapy, a radioisotope as well as their major subtypes, or a combination of modalities.

3.2.6. Additional Non-Morphological Variables

The Surgery variable included no surgery, autopsy, the patient died before recommended surgery, unknown reason for surgery, the patient died after surgery, and surgery to the distant lymph node. The age at diagnosis represents the patient's actual age in years. The number of the lymph node is a time-dependent prognostic factor. The prognosis worsens if the number of nodes involved increases (www.seer.cancer.gov).

3.3. Notations

First and foremost, we will let x_{ij} represent the value of the jth variable for the ith observation, where $i = 1,2, \ldots \ldots \ldots, n$ and $j = 1,2, \ldots \ldots \ldots, p$. Throughout this dissertation, i will be used to index the patients (from 1 to n) and j will be used to index the predictor variables (from 1 to p).

Here, x_i is a vector of length p, containing the p variable measurements for the ith patients That is,

$$x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix} \quad (1)$$

We use y to denote the response variable, i.e., survival status, of i^{th} patients and is the variable on which we wish to make cancer-specific survival predictions. We can write the set of all n responses in vector form as:

$$\text{y}= \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad (2)$$

Our data consists of $\{(x_{i1}, y_1),\ (x_{i2}, y_2)\ \ldots\ldots\ldots\ldots (x_{ip}, y_n)\}$, where each $\boldsymbol{x_i}$ is a vector of length of p.

3.4. Data Normalization

Normalizing, or transforming, data before performing PCA is a common approach when variables are measured on different scales or have a wide range of variances. The main goal of this transformation is to make the variances of the variables comparable so that variables with large relative variances do not dominate the top principal components. Two common methods of transforming variable measurements include Min-Max normalization and Z-score standardizations In the Min-Max normalization approach, for each variable, the maximum measurement gets transformed into a 1, the minimum measurement gets transformed into a 0, and each other value gets transformed into a value between 1 and 0:

where $min(X_j)$ and $max(X_j)$ are the minimum and maximum, respectively, of the x_{ij} values ($i = 1,2, \ldots, n$).

In the Z-score standardization approach, for each variable, the mean of the measurements is subtracted from each measurement, and the differences are divided by the standard deviation of the measurements:

$$x_{ij}^{*} = \frac{x_{ij} - \min(X_j)}{max(X_j) - \min(X_j)} \tag{3}$$

$$\text{Z-score: } x_{ij}^{*} = \frac{x_{ij} - mean(X_j)}{sd\,(X_j)} \tag{4}$$

where $mean(X_j)$ and $sd(X_j)$ are the mean and standard deviation, respectively, of the x_{ij} values $(i = 1,2,\dots,n)$. This approach results in a normalized data set where each variable has measurements with a mean of 0 and a standard deviation of 1.

3.5. One Hot Encoding Method

One hot encoding is very useful for categorical variables with more than two classes. To implement this process, the encoded variable is removed, and a binary variable is created for each unique class by assigning a binary value of 0 or 1 to those columns. Therefore, each value is indicated by a binary variable. Table 3 and Table 4 illustrate the process of one hot encoding method.

Table 3. Example of categorical value using car company

Company Name (Class)	Categorical value	Price
Toyota	1	20000
Acura	2	10011
Honda	3	50000
Majda	4	35000

In Table 3, the categorical value indicates the class of the entry in the dataset. In Table 4, each class is represented by a binary variable.

Table 4. One hot binary encoding

Company Name	Toyota	Acura	Honda	MajdA
Toyota	1	0	0	0
Acura	0	1	0	0
Honda	0	0	1	0
Majda	0	0	0	1

We use one-hot encoding to perform "binarization" of the categorical variables and include them as features to train the model.

3.6. Data Reduction and Variables Selection Techniques

Variable selection is the process of identifying important variables and removing redundant features in a dataset. This is a useful technique in many predictive and statistical problems because of the high dimensionality of the predictor variables. There are two methods of variable selection the Lasso, and Boruta algorithms. PCA reduces dimensionality.

3.7. Principal Component Analysis (PCA)

Principal components are uncorrelated linear combinations of the p variables from the observations in a sample that account for the maximum sample variance. The total number of principal components is equal to the number of original variables. However, principal component analysis is often performed to reduce the dimensionality of a large dataset that includes many variables into a new dataset that includes a much smaller number of principal components while still accounting for a high proportion of the original total sample variance, making it easier to

analyze and visualize the data with minimal loss of information. In this new data set, the first principal component is the linear combination of the original variables accounting for the maximum total sample variance. The second principal component is a linear combination of the original variables that are orthogonal, or uncorrelated, with the first principal component that accounts for the maximum remaining total sample variance. This process can be repeated until *p* principal components are obtained, although as previously mentioned, a small number of principal components is commonly used in subsequent analysis. The k^{th} principal component, PC_k is

$$PC_k = a_i'x = a_{i1}x_1 + a_{i2}x_2 + \ldots\ldots + a_k p_{xk} \quad (5)$$

Where ak_j is the j^{th} element of the k^{th} eigenvector of **S**, the sample covariance matrix of *n* patients vectors, $x_{i1}, x_{i2}, \ldots\ldots\ldots, x_{ip}$

In our analysis, we used the functions prcomp() and PCA() for the singular value decomposition (SVD) approach to calculating eigenvectors. before generating the principal components, The summary() function is used to get the percentage of variance explained in the predictors.

3.8. Boruta Algorithm Techniques

There are a lot of reasons to use the Boruta algorithm for feature selection, including the following.

- It can be used for both regression and classification problems.
- Its variable importance measure is an improvement over that of the random forest method because the random forest method only uses the Mean Decrease Accuracy (MDA) or Mean Decrease Gini (MDG) to evaluate the importance of each variable. On the other hand, the Boruta algorithm follows other additional steps including MDA and MDG to get the significant variables.

- It considers the correlations and interactions between the variables.
- It is a very strong tool for removing redundant features and retaining features that are relevant to the outcome variable (Kursa & Rudnicki, 2010).

Boruta algorithm can be implemented on datasets when one is interested in understanding the variable of interest for prediction accuracy. In our dataset, rows correspond to patients and columns correspond to predictor variables. The following steps are as follows for the Boruta algorithm:

Steps of Boruta algorithms

1. The predictor variables are duplicated and the values in each column are shuffled to remove their correlations with the target variable. These shuffled values are called shadow features or permuted copies.
2. The shuffled copies are combined with the original feature.
3. The Mean Decrease Accuracy or Mean decrease Gini are computed and used to evaluate the importance of each variable:

   Mean Decrease Accuracy = $\frac{1}{ntree} \sum_{t=1}^{ntree} EP_t - E_t$

where tree indicates the number of trees in the forest, EP_t denotes the out-of-bag error on tree t before permuting, and E_t denotes the out-of-bag error on tree t after permuting (Han et al., 2016). The out-of-bag error is a method of measuring the prediction error by evaluating predictions on those observations that were not used in the training set.

$$Mean\ Decrease\ Gin = \frac{\text{The total decrease in node impurities from splitting on the variable}}{Number\ of\ trees}$$

Node impurity indicates the measure of the homogeneity of the labels at the node. The homogeneity measured by Gini index, entropy, and information gain. Containing similar values with contain instances into the subsets of data is called homogeneity.

4. The Z score is computed. In the context of the Boruta algorithm, the Z score is computed by dividing the mean decrease in accuracy by its standard deviation. It is used as an important measure.
5. The maximum Z score is determined among shadow attributes The Z score of the original features and the shuffle copies are compared at every iteration. If the original feature has the maximum Z score, then this feature is tagged as important. Otherwise, it is considered unimportant.

3.9. Lasso Regression Model

We will first illustrate the linear model before illustrating the Lasso model.

The linear model defined as

$$f(x) = \beta_0 + \sum_{j=1}^{p} x_j \beta_j \tag{6}$$

β_0 is intercept and unknown parameter. β_j is also unknown parameter and slope coefficients.

$$RSS\ (\beta) = \sum_{i=1}^{N} (y_i - f(x))^2$$

$$RSS\ (\beta) = \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_j \beta_j)^2 \tag{7}$$

From the above equation, the Ordinary Least Square (OLS) approach used to choose $\widehat{\beta_0}$ and $\widehat{\beta_J}$ to minimize the RSS.

Lasso regression is widely used for feature or variable selection and shrinks the regression coefficients by imposing a penalty on their size. In 1986, (Santosa & Symes, 1986) independently developed the Lasso regression model. In 1996, Statistician Robert Tibshirani independently improved the Lasso model. Lasso regression improves prediction error and performs covariate selection by forcing the sum of the absolute value of the regression coefficients to be less than a fixed value, which forces some coefficients to be exactly zero (Tibshirani, 2011). Only the most significant variables are considered in the final model. A penalty term is added to the log-likelihood function of regular regression due to lasso regularization. The Lasso coefficients minimize a penalized residual sum of the square. The lasso estimates (Gareth et.al., 2013) are defined by:

$$\widehat{\beta lasso} \;=\; RSS\,(\beta) = \; argmin\left\{\frac{1}{2n}\sum_{t=1}^{n}(y_i - x_i\;\beta)^2 \;+\lambda\sum_{j=1}^{p}\left|\beta_j\right|\right\}, \qquad (8)$$

where, the λ is considered as a penalty parameter, y_i is the outcome variable, β is the vector of coefficients on x, β_j is the jth element of β, n is the number of patients.

There are two terms in this optimization problem. The first is the least-squares fit measure,

$$\frac{1}{2n}\sum_{t=1}^{n}(y_i - x_i\;\beta)^2.$$

The second is the penalty term: $\lambda\sum_{j=1}^{p}\left|\beta_j\right|$.

The parameter λ is called the "tuning" parameter. When $\lambda = 0$, the linear lasso estimators reduce to the OLS estimators. They specify the weight applied to the penalty term. The magnitudes of all the estimated coefficients are "shrunk" toward zero as λ increases.

We used the glmnet R package, first introduced by Friedman et al. (2010), for lasso feature selection (Hastie & Qian, 2014). The main function in this package is glmnet(), which can be used to fit lasso models. The model. matrix () function is particularly useful for creating the design matrix; not only does it produce a matrix corresponding to the predictors but also transforms any qualitative variables into dummy variables. The latter property is important because glmnet() can only take numerical, quantitative inputs.

One important function included in the glmnet package is cv.glmnet (). This function finds the optimal value of λ, defined as the λ that minimizes the cross-validation prediction error rate. This function performs cross-validation (CV) by 10-fold which can be adjusted using the number of folds. Overall, the main point is that the lasso selects significant features by shrinking the coefficients of unimportant features to zero.

3.10. Cancer Survival Prediction Methods

There following seven machine learning techniques are used to predict cancer survival: LR, RF, DT, ANN, LDA, NB, and SVM. We discussed every method in detail in the following sections.

3.11. Decision Tree

The decision tree is a non-parametric supervised learning method introduced by Breiman (1984) that uses decision rules to predict the value of a quantitative variable or perform binary classification. Each node is determined based on the entropy.

Figure 1 shows the root node, decision nodes, and terminal or leaf nodes for the decision tree.

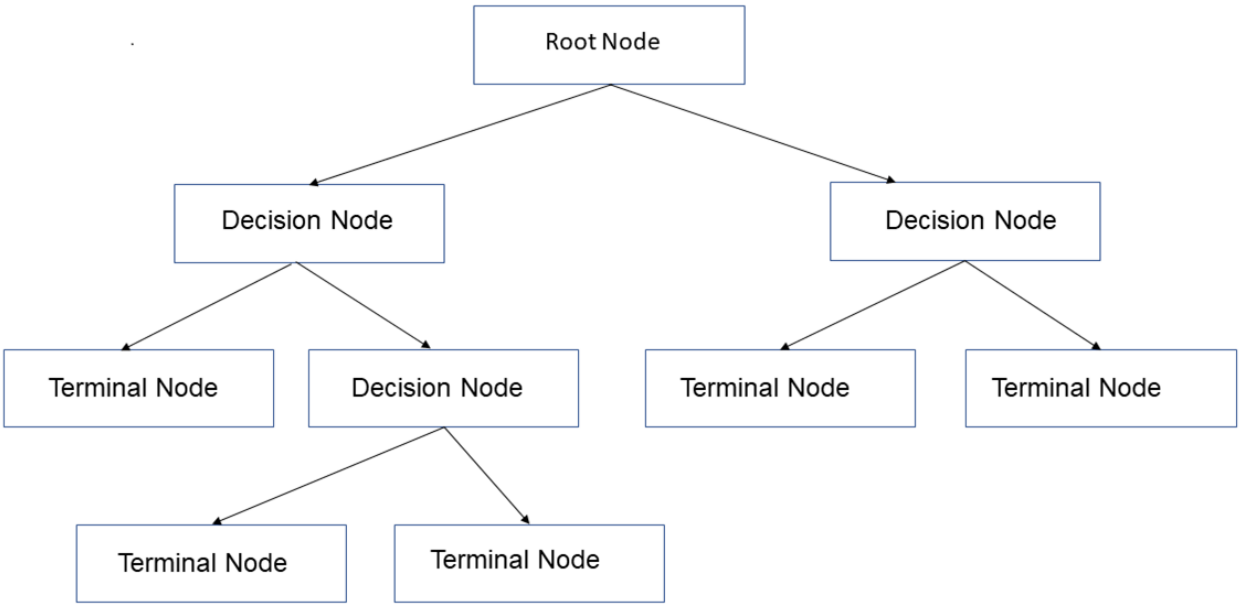

Figure 1. Different nodes of decision tree

The root node indicates the whole sample or population. It is further split into two or more homogenous sets. If this sub-node is divided into further sub-nodes, this is called a decision node. A terminal node or leaf node is a node that does not split (Figure 1).

Here, we present how decision trees are used for binary classification. The decision tree process divides the predictor space indicated by $X_1, X_2, \ldots\ldots\ldots X_p$ into K distinct regions $R_1, R_2, \ldots\ldots\ldots R_K$. Then, the top-down greedy approach, also known as recursive binary splitting, is performed. This approach starts at the top of the tree and subsequently splits the predictor space. Every split is created via new branches further down on the tree as follows. Let the predictors' space be split into the region by cutting point *t*. The regions $\{X|X_j < t\}$ and $\{X|X_j \geq t\}$ lead to reduced classification error for the classification tree and Residual Sum of Square (RSS) for the regression trees. The cutting point t is determined by evaluating the entropy for each variable. The notation $\{X|X_j < t\}$ indicates the region of predictor space in which X_j takes on a value less than *t*. The notation $\{X|X_j \geq t\}$ means the region of predictor space in which

X_j takes on a value greater than or equal to t. We define the pair of half-planes for two regions for any j and t as

$$R_1(j,t) = \{X|X_j < t\} \; and \; R_2(j,t) = \{X|X_j \geq t\} \tag{9}$$

The response for a given test observation is predicted using the mean of the training observations in the region to which that test observation belongs once the regions $R_1, \ldots, R_k$ have been created.

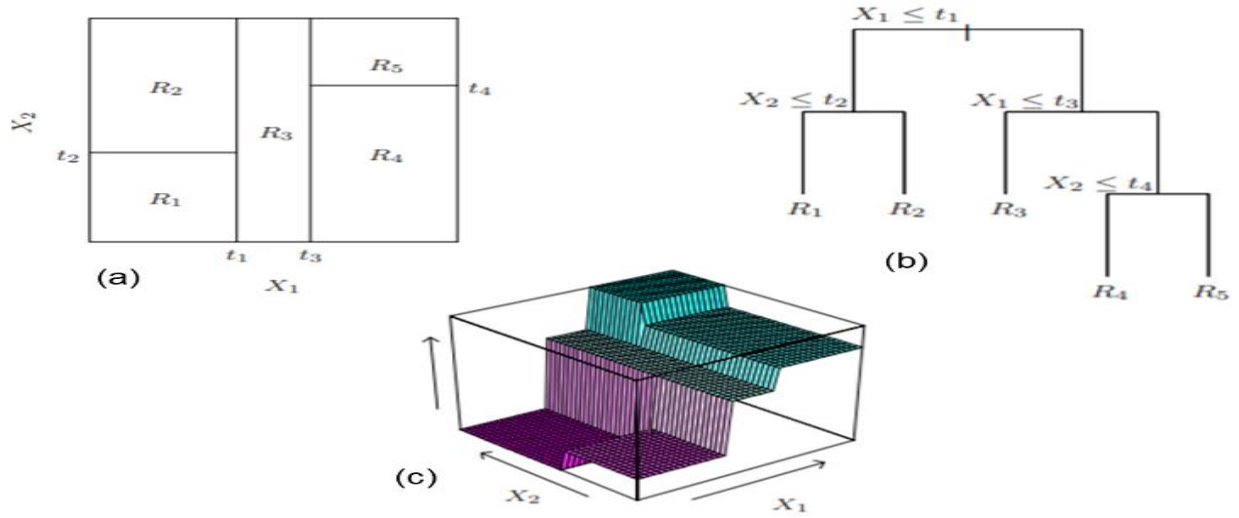

Figure 2. How the decision trees work

Figure 2 shows an example of decision rules for a decision tree. In the top left panel, first, we split at $X_1 = t_1$. Then the region $X_1 \leq t_1$ is split at $X_2 = t_2$ and the region $X_1 > t_1$ is split at $X_1 = t_3$. Finally, the region $X_1 > t_3$ is split at $X_2 = t_4$. The result of this process is a partition into the five regions $R_1, R_2, \ldots, R_5$ shown in Figure 1. The regions of $R_1, R_2, \ldots, R_5$ corresponds to the terminal nodes or leaf node. Both panels (a) and (b) illustrate the same decision process if $X_1 \leq t_1$ and $X_2 \leq t_2$, the region would be considered R_1 and R_2. In addition, if $X_1 \geq t_1$ and $X_1 \leq t_3$ the region or terminal would be R_3. On the flip side, if $X_1 \geq t_1$, $X_2 \geq t_3$, and $X_2 \leq t_4$ the region covers R_4, otherwise R_5. The bottom panel shows a perspective plot of the prediction surface corresponding to that tree (Gareth et.al., 2013).

The classification tree is quite like the regression tree. In the classification tree, the classification error rate is considered for making the binary split instead of RSS. The classification error rate, ERR, is defined as the proportion of patients classified in the wrong class.

$$\text{ERR} = \frac{Number\ of\ classified\ \ wrong\ patiens}{Total\ number\ of\ patients}$$

Because the classification error is not sufficiently sensitive for classification trees, the Gini Index and cross-entropy measures are preferable for node impurity. The Gini Index and entropy are error metric that is designed to show how "pure" a region is. "Purity" in this case means how much of the training data in a particular region belongs to a single class. We used a classification error rate to get our prediction accuracy.

The Gini index is defined by

$$GINI = 1 - \sum_i [Y\left(\frac{i}{t}\right)]^2 \tag{10}$$

$Y\left(\frac{i}{t}\right)$is the relative frequency of class I at node t. The Gini Index is a node impurity measure that indicates how much of the training data in a particular region belongs to a single class. If a region contains data that is mostly from single class I, the Gini index value will be small. A small Gini index is desirable for the purity of the node. Another measurement is cross-entropy (Gareth et.al., 2013), given by

$$Entropy = -\sum_i Y\left(\frac{i}{t}\right)\ log\ Y\left(\frac{i}{t}\right) \tag{11}$$

Tuning parameter for decision tree: Node impurity is the tuning parameter for the decision tree. It is determined by entropy and the Gini index. Our goal was to identify the variables which have the highest amount of information, conversely the lowest entropy. Based on this information, the first split towards either the left or the right of the root node. Then we repeated this process

under the root node and calculated all possible splits and the split chosen with the lowest Gini score. The process repeated for both left and right sides until we reached terminating a class in the response variable. After that, we fitted our model to get a confusion matrix for model performance. For the whole process, we used the “rpart” package in R.

The “rpart” R package was used for decision tree analysis. First, we split the observations into a training set and a testing set, built the tree using the training set, and evaluated its performance on the testing data. The predict() function was used to evaluate model performance. The argument type="class" instructs the class label survived, and not survived.

3.12. Random Forest

The random forest algorithm is a machine learning technique introduced first by Tin Kam Ho in 1995 (Ho, 1995), and later modified by Adele Cuter and Leo Breiman in 2001(Breiman, 2001). Random forest is a very popular technique for the modification of bagging. Bagging is a technique used to reduce the variance of prediction by combining the results of multiple decision trees on different sub-samples on the same dataset. The steps to construct the random forest are as follows.

Steps of random forest algorithm:

1. Draw a bootstrap sample of size d from the training data.
2. Draw a random forest tree to the bootstrapped data by recursively repeating the following steps for each terminal node of the tree until the minimum node size is reached.
    i. Select *p* variables at random from the training data.
    ii. Pick the best split point from the *p* variables as determined by the entropy.
    iii. Split the node into decision nodes (Hastie et al., 2009)

3. Output the from the bagging and bootstrap aggregation is used to predict a new point *t* for the classification tree. Let $\hat{Y}b(t)$ be the class prediction of the b[th] random forest tree.

$$Y_{rf}^{B}(t) = majority\ vote\ \{\hat{Y}b(t)\}_{0}^{B} \quad (12)$$

The Majority vote indicates the majority decision tree gives output survival class compared to the non-survival class. voting was performed for every predicted result. Therefore, select the most voted prediction result as the final prediction result.

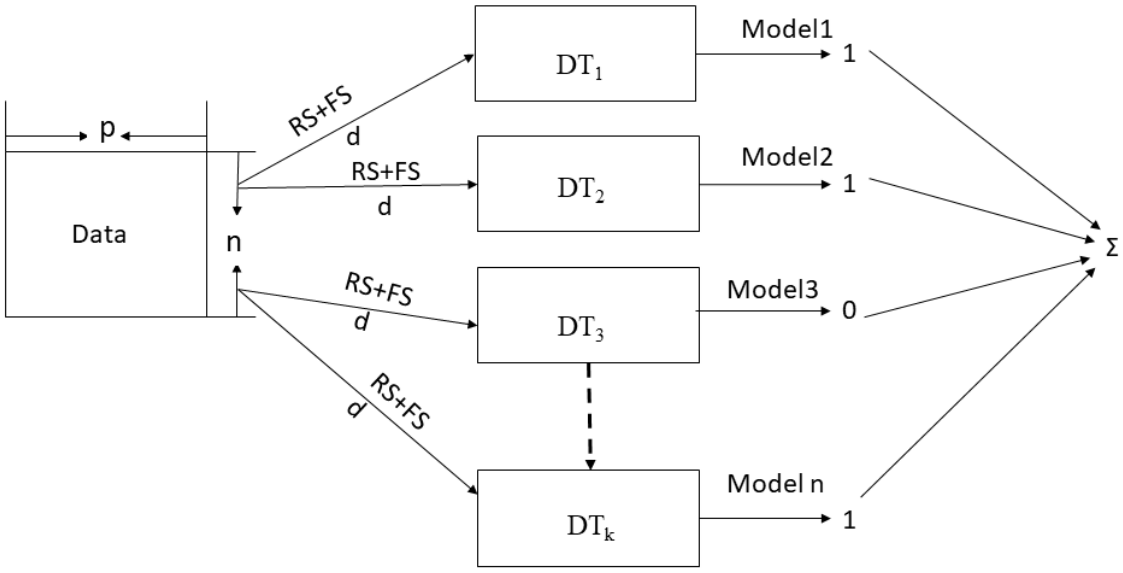

Figure 3. Random forest algorithm ((Tan et al., 2016)

Here, *p* is the number of columns (variables) in the data set, *n* is the number of records (observations) in the data set, d is a random sample from n RS represents row sampling, FS represents feature sampling, and DT represents a decision tree (Figure 3).

Random forest gives more accuracy than the single decision tree and identifies which variables are important in the classification. The variance in the final prediction is reduced by averaging the predictions of the multiple random forest trees, therefore improving the predictive performance over a single decision tree. The random forest also ranks the importance of variables in classification through Mean Decrease Accuracy (MDA) and Mean Decrease Gini (MDG).

MDA is a method of computing the variable's importance on permuted out-of-bag samples based on the mean decrease in accuracy. MDG is a measure of variable importance based on the Gini impurity index utilized for the calculation of splits in trees. Indeed, the node impurity is measured by the Gini index (Genuer et al., 2010). The higher the value of mean decrease Gini or mean decrease accuracy, the higher the importance of the variable in the model (Breiman, L. 2001).

We applied bagging and random forests using the randomForest package in R. It was developed by Breiman in 2001. Bagging is simply a special case of a random forest in which *p* variables are selected with replacement. We can view the importance of each variable using the importance () function. Using the varImpPlot() function, the plots of each variable importance can be produced.

Two parameters that are important in the random forest function ntree and mtry are the number of trees used in the forest (ntree) and the number of random variables used in each tree (mtry). First, we set mtry to the default value (the square root of the total number of all predictors) and search for the optimal ntree value. To find the number of trees that correspond to a stable classifier, we built a random forest with different ntree values. We built random forest classifiers for each ntree value and observe the number of trees where the out-of-bag error stabilizes and reaches a minimum.

3.13. Artificial Neural Network

McCulloch and Pitts first introduced the mathematical model of a neuron in 1943 in their paper, “A logical calculus of the ideas immanent in nervous activity”. In their research paper, they explained the simple mathematical model for a neuron that performs similarly to a single cell of the neural system that takes inputs, processes those inputs, and returns an output (McCulloch & Pitts, 1943). In 1969, Minsky and Papert ascertained two issues with neural networks. The first

issue was that single-layer neural networks were not capable of processing the circuit and another issue was computers were not sophisticated enough to handle large neural networks (Minsky & Papert, 1969). Rumelhart et.al. (1986) presented a full explanation of the connectionism in computers to simulate the neural process. They also developed a back-propagation algorithm which was the most popular algorithm for multilayer perceptron. A perceptron is an algorithm for supervised learning of binary classifiers. A binary classifier is a function that can decide whether an input, presented by a vector of numbers, belongs to some specific class. There are two types of perceptron, single layer, and multilayer. Multilayer perceptron indicates two or more hidden layers.

Neural Networks are inspired by the human brain to perform a function or particular task. It is now a tool of artificial intelligence and is widely used in machine learning. An artificial Neural Network (ANN) is treated as a non-linear modeling technique in statistics. It is a very powerful prediction tool that identifies complex patterns within a dataset. There are k layers of interconnected units in a neural network called neurons. In this study k =2 hidden layers are used for illustration. The hidden layer is directly connected with each unit of the input layer. (Yeh, 2019) defined the neural network as a classified triple (N, V, ω) with a function ω and two sets N and V, where V is the set $\{(i, j)|i, j \in N\}$ and N is the set of neurons. The weighted sum is the most popular function which, for a given neuron j, is given by

$$W_i = \sum_{=1}^{n} f_i * w_i \tag{13}$$

where n is the number of neurons in the previous layer, f_i is the output of the previous layer for the i^{th} neuron, w_i is the weight that indicates the contribution of input x_i to the perceptron output. Weight is the parameter within a neural network that transforms input data within the network's hidden layers and can be the strength of the connection between input and hidden layers. A small

weight causes little change in the input, and a larger weight value causes a significant change in the output. A neural network has a series of nodes that indicates neurons. Figure 4 shows the single-layer neural network. It has three neurons (f_1, f_2, and f_3) for one hidden layer. The second function is called the activation function or output function.

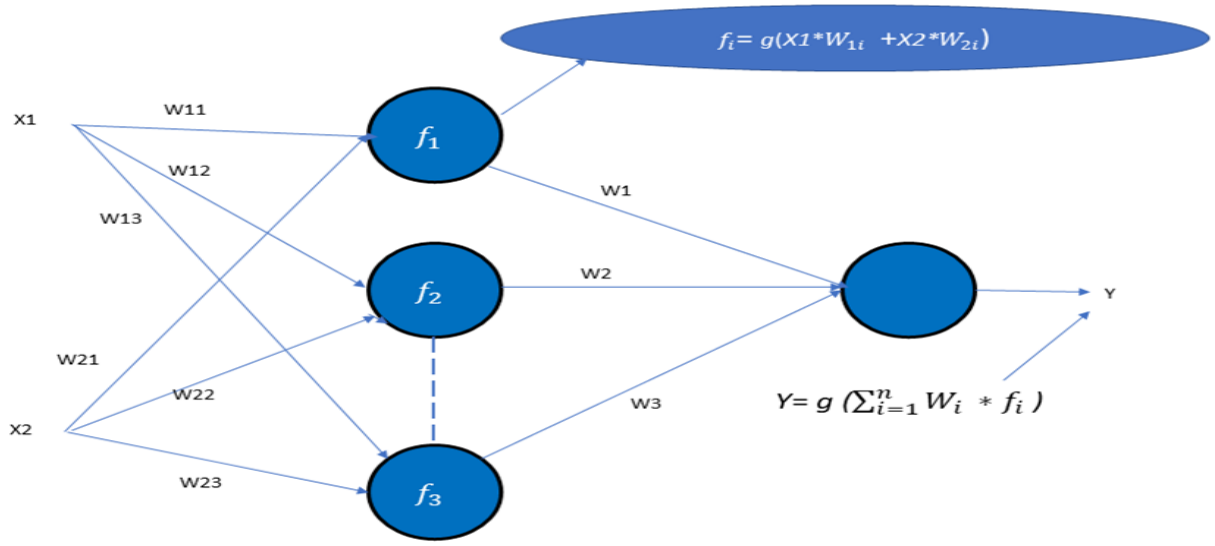

Figure 4. One Possible Structure for Artificial Neural Network

From Figure 4, $f_i = g(x_i)$, i=1,……n are activation functions. An activation function is assigned to the entire layer of perceptron and the weighted sum of input values is added up. There are several functions used as activation functions such as linear functions, sigmoid, and binary step functions. The sigmoid function is a logistic function where the output values vary from 0 to 1. In our study, we used the sigmoid function because the output values are binary. The nonlinear sigmoid function is given by

$$\varphi_{logistic\ Y} = \frac{1}{1 + exp^{-Y}} \tag{14}$$

where the value of $Y = W_0X_0 + W_1X_1 + \ldots\ldots + W_iX_i = \sum_0^i W_i\ X_i = W^T\ X$

Algorithms of Neural Networks: The following algorithm steps are used for neural network

- Step 1: First derived features for hidden units including activation function.

- Step 2: Compute a linear combination of the features for hidden units
- Step 3: Compute the output function with weights. (Gareth et.al., 2013).

Parameters tune for ANN: Hyperparameter tuning is important for ANN analysis. There is no specific answer, how many layers are most suitable, and how many neurons are the best? But hyperparameter tuning is to find the best possible hyperparameter to build the model. We considered the number of neurons, activation function, and the number of layers as hyperparameters for tuning. In our analysis, two hidden layers were used with 10 neurons because we observed that prediction accuracy bit improved instead of one hidden layer with different neurons.

We used neuralnet R package for multilayer neurons. In 1994, Riedmiller developed a backpropagation neuralnet package for single-layer neurons. This package was modified by Anastasiadis et al. in 2005 for multilayer neurons, error, and activation function. We used various options for analysis from neuralnet package: formula, hidden, threshold, step max, err. fct, linear. output. The hidden argument accepts a vector specifying the number of neurons in each hidden layer. The threshold argument indicates a numeric value specifying the partial derivatives of the error function as a stopping criterion. The error function is the sum of squared of the differences between expected output and actual output. Stepmax is a very important option for running the maximum steps of the neural network. It leads to a stop of the neural network process. Linear. output is used to specify whether we want to do classification or regression.

3.14. Support Vector Machine (SVM)

The Support Vector Machine (SVM) method was originally proposed by Vladimir and Alexey in 1963. In 1992, a nonlinear classifier was presented by Bernhard and Isabelle Guyon (Boser et al., 1992). The SVM improvement was proposed by Corinna in 1995 (Cortes & Vapnik,

1995). After Corinna, the SVM was further developed by Vladimir in 1998. SVM is a supervised machine learning technique used for regression and can handle both non-linear and linear class boundaries. The kernel function is used to transform the data into a higher dimensional feature space and constructs the decision boundary, called a hyperplane, that best separates the classes. We considered radial, linear, and sigmoid kernels in our analysis. Figure 5 shows the linear and nonlinear boundaries.

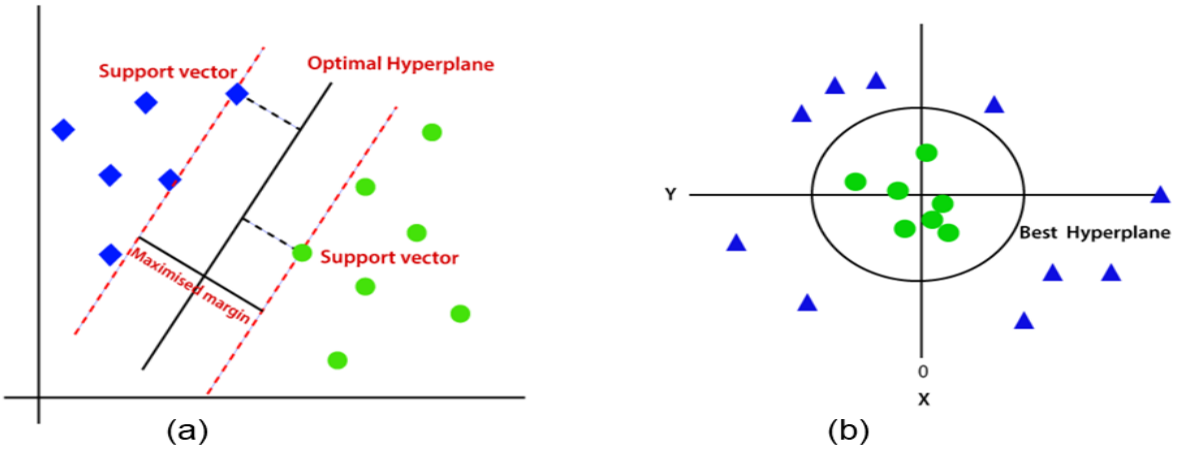

Figure 5. How SVM Algorithm work (https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm)

The figure in the top left panel shows linear SVM. If a dataset can be classified into two classes (blue point and green point) by using a single straight line or hyperplane, this is called linearly separable data. The nearest data points of the hyperplane from both the classes are called support vectors. Two blue data points are very close to the hyperplane. Therefore, the marginal plane passes through the two support vectors. On the other hand, one green data point is very nearest to the hyperplane. Consequently, the marginal plane passes through one support-vector. The distance between the hyperplane and the vectors is called the margin**.** The goal of SVM is to maximize this margin. The hyperplane with the maximum margin is called the optimal hyperplane**.**

The top right panel of Figure 5 shows the nonlinear SVM. If a dataset cannot be classified by using a straight line, this is called nonlinear SVM. A third dimension is needed for non-linear data. Therefore, we get a circumference of radius 1 in the case of non-linear data.

Steps of algorithms for SVM:

1. First and foremost, the SVM algorithm finds the best line
2. SVM algorithm finds the closest point of the lines
3. The SVM algorithm is to maximize the margin.
4. Transform data to high dimensional space where it is classified with linear decision surface. In this step, kernel function involved because data transform depends on the linear, radial, and sigmoid kernel function. Since our data classified into two parts survival and not survival, therefore, we considered linear decision surface.
5. Finally, maximizing the margin is to get an optimal hyperplane.

There are some kernel functions used in SVM analysis such as linear, radial, polynomial, and sigmoid kernel functions. Th**is** is a method used to take data as input and transform it into the required form of processing data. The polynomial kernel function computes the degree-d polynomial kernel between two vectors. and represents the similarity between two vectors. The polynomial kernel is also considered across dimensions between two vectors with the function. The polynomial kernel function is as follows:

$$K\,(a,b)\;=\;(\gamma a^{T}b+\;C_{0})^{d} \tag{15}$$

where a and b are the input vectors, d is the kernel degree, γ is the slope and C_0 is the intercept

The linear kernel function is

$$K\,(a,b)\;=\;\;a^{T}\,b \tag{16}$$

for observation vectors a and b.

The radial kernel function is

$$K(a,b) = \exp\left(-\frac{||a-b||^2}{2\sigma^2}\right) \quad (17)$$

where $||a\text{-}b||^2$ indicates the squared Euclidean distance, σ is a free parameter and the kernel is known as the Gaussian kernel of variance of σ^2. The sigmoid kernel is also known as a hyperbolic tangent or Multilayer Perceptron and has the function

$$K(a,b) = tanh\left((\gamma a^T b + C_0\right) \quad (18)$$

where a and b are the input vectors, γ is the slope, and C_0 is the intercept.

We used the e1071 library in R to implement the support vector classifier and the SVM. When the argument kernel="linear" is used, the SVM() function can be used to fit an SVM() support vector classifier. A cost argument allows us to identify the cost of a violation to the margin. The margins are wide and many support vectors are based on the margin or violated margin if the cost argument is small.. We obtained basic information about the support vector classifier fit using the summary () command. We used kernel="polynomial" to fit an SVM with a polynomial kernel and to fit an SVM with a radial kernel we used kernel="radial".

The plot() function is used to represent data, models, and support vectors in a visual form. It can also be used to build a model with a scatter plot of input. The predict() function is used to predict the classes of the test set observations.

3.15. Discriminant Analysis (DA)

Based on one or more independent or predictor variables, discriminant analysis is used to predict the probability of an observation belonging to a given category or class. It works with both categorical and continuous independent variables. We consider the linear discriminant analysis in our study. Linear discriminant analysis is used to predict the class of a given observation through

linear combinations of predictors. It was first developed by Ronald Fisher in 1936 to classify individuals into one of two clearly defined groups (Fisher, 1936). The class posterior probabilities *Pr(y|x)* for optimal classification are important to know according to the decision theory for classification. Suppose $f_k(x) = Pr\ (x|y(x)=k)$ is the class-conditional density of observation x in class $y = k$, and $\pi_k = Pr\ (x\ from\ group\ k)$ is the prior probability with

$$\sum_{k=1}^{K} \pi_k = 1 \tag{19}$$

where, k ϵ {1,2,……, K}. Bayes theorem then gives us the posterior probability that an observation belongs to group *k:*

$$Pr(y(x) = k|x)\ \frac{Pr\ (k,x)}{\Pr\ (x)} = \frac{\pi_k f_k\ (x)}{\sum_{k=1}^{k} \pi_k f_k\ (x)} \tag{20}$$

.The decision rule for classifying a new observation y_0 is to assign y_0 to class k_0 if

$$k_0 = argmax k\ \epsilon\ \{1,2,\ldots.,K\}\ Pr\ (k|y_0) \tag{21}$$

Here, k_0 is the group label such that $Pr\ (k_0|y_0)$ is the largest among

$$Pr\ (k|y_0), k\ \epsilon\ \{1,2,\ldots\ldots,K\} \tag{22}$$

The prior probability π_k is estimated as

$$\pi_k = \frac{N_k}{N}$$

where N_k is the number of observations in the training data set below to group *k.*

The densities $f_k(x)$ can be determined or estimated using the following techniques:

- linear discriminant analysis uses Gaussian densities
- more flexible mixtures of Gaussians can be used to allow for nonlinear decision boundaries

- general nonparametric density estimates for each class density allow the most flexibility
- Naive Bayes models assume that each of the class densities is the product of marginal densities. In addition, the inputs are assumed to be conditionally independent in each class.

Linear Discriminant Analysis (LDA) assumes $f_k(x)$ is Gaussian with equal covariance structure among the K groups (Gareth et.al., 2013). The density for class k is

$$f_k(x) = \frac{e^{-\frac{1}{2}(x-\mu_k)^T \Sigma^{-1}(x-\mu_k)}}{2\pi^{p/2} \ |\Sigma_k|^{1/2}} \tag{23}$$

where f_k is the class conditional density of x in class $y=k$, μ_k is the length of a p row vector, x is a vector of p values, $x-\mu_k$ is a row vector, Σ^{-1} is $p*p$ matrix, $(x-\mu_k)^T$ is a column vector, $2\pi^{p/2} \ |\Sigma_k|^{1/2}$ is constant.

Suppose $\sum_k = \Sigma, k = 1,2, \dots \dots, K$

$$= log\frac{\pi_k}{\pi_l} - \frac{1}{2}(\mu_k + \mu_l)^T \Sigma^{-1}(\mu_k - \mu_l) + X^T \Sigma^{-1}(\mu_k - \mu_l)$$

The linear discriminant function is defined as

$$\delta_k(y) = X^T \ \Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T \Sigma^{-1}\mu k + log\,\pi_k . \tag{24}$$

The following properties are important for LDA

- LDA assumes the same covariance matrix for all classes.
- LDA is not suitable if there are higher-order interactions between predictor variables. For k-1 dimensional subspace, it finds linear boundaries.
- LDA can be used for classification prediction and dimensionality reduction (Trevor et.al., 2013).

We used the MASS package in R for performing linear discriminant analysis. The LDA algorithm was applied to find a discriminant function that can maximize the separation among the classes. Then it uses these discriminant functions for predicting the class of every individual. This discriminant function is called linear discriminant function and is a linear combination of the explanatory variables**.** The lda () function was used in the MASS package to specify the prior probabilities of groups, group means, and co-efficient of the linear discriminant. In addition, the cross-validation (CV) method was implemented by lda() function. Furthermore**,** lda() shows the mean of each variable in each group and prior probabilities. Finally, the linear combination of predictor variables was used to form the LDA decision rule. The LDA decision rule assumes the equality of covariances the predictor covariates x across all classes.

3.16. Naïve Bayes Classifier

Naïve Bayes classification is called a Bayesian classifier which is a probabilistic model based on Bayes' theorem. It can predict the class probabilities such as the probability that the given i predictor variables. The maximum probability is called a maximum posterior hypothesis, and this can be calculated by using the Bayes theorem. Using Bayes theorem:

$$Posterior\ probability = \frac{Class\ Prior\ probability * likelihood}{Evidence}$$

The above can be written as follows:

$$P(y|x) = \frac{P(x|y)*p(y)}{P(x)}. \quad (25)$$

- *P(y)* represents the prior probability of the y
- *P(x)* represents the prior probability of evidence
- *P(y|x)* is called the posterior
- *P(x|y)* is the likelihood (Trevor et.al., 2013)

Hence, we reach to the results

$$P(y|x_1, \dots \dots \dots x_p) = \frac{P(x_1|y)P(x_2|y) \dots \dots \dots \dots P(x_p|y)P(y)}{P(x_1)P(x_2) \dots \dots \dots P(x_p)}$$

It can be expressed as:

$$(y|x_1, \dots \dots \dots x_p) = \frac{P(y)\prod_{j=1}^{P} P(x_j|y)}{P(x_1)P(x_2) \dots \dots P(x_p)}$$

As the denominator remains constant for a given input, we can remove that term:

$$P(y|x_1, \dots \dots \dots x_p) \propto P(y)\prod\prod_{j=1}^{p} P(x_j|y)$$

We can find the probability of given set of inputs for all possible values of the class variable *y* and pick up the output with maximum probability. This can be expressed mathematically as:

$$y=\text{argmax}P(y)\prod_{j=1}^{p} P(x_j|y)$$

We used the naivebayes package to implement the naïve bayes classifier. The function naive_bayes() detects the class of each feature in the dataset and computed prior and posterior probabilities

3.17. Logistic Regression

Logistic regression is a statistical model which has the basic form of the logistic function to model a binary response variable. A dependent variable with two possible values, denoted by 1 and 0, is considered in the logistic model. Predictors can be continuous or binary variables. Traditional binary logistic regression is based on Maximum Likelihood Estimation. The coefficient estimates maximize the likelihood function. The function of logistic regression (Gareth et.al., 2013) is as follows

$$log(\frac{P(y)}{1-P(y)}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k \quad (27)$$

The above equation on the right side expressed the log odds. *P(y)* indicates the probability of survival. . *1-P(y)* is the probability of non-survival. On the other hand, the right side expressed the outcome is linked to the linear predictors. β_i is the slope of log odds.

We can recover by exponentiating the log odds as

$$\frac{P(y)}{1-P(y)} = e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k} . \quad (28)$$

The above equation can be re-expressed as

$$P(y) = \frac{e^{\beta_0 + \beta_1 X_1 + \cdots\ldots + \beta_k X_k}}{1+e^{\beta_0 + \beta_1 X_1 + \cdots\ldots + \beta_k X_k}}. \quad (29)$$

where *P (y)* is the probability that the dependent variable y, given a linear combination of the predictors. The probability *P(y)* ranges between 0 and 1, $\frac{Py}{1-P(y)}$ is the ratio of the probability of survival to the probability of non-survival, which is called the odds ratio, β_0 and β_i are parameters to be estimated, and p is the number of predictors (Hosmer et al., 2000). Logistic regression in R is implemented using the "glm" function. This glm() function is used to fit generalized linear models, including logistic regression models Binomial distribution is used to fit a logistic regression. A GLM (Generalized Linear Model) uses the log link function because the response variable is categorical.

3.18. K- Nearest Neighbors (KNN)

The K-Nearest Neighbors (KNN) method was first introduced by Joseph Hodges and Evelyn Fix in 1951 and was expanded by Thomas Cover in 1967 (Cover & Hart, 1967). KNN is a non-parametric supervised learning technique. With the help of a training set, data points are classified into a given category. Specifically, the classification of new cases is based on their

closeness to neighbors. The steps of the KNN algorithms for classifying each observation are as follows:

1. Select K, the number of the neighbors. Therefore, calculate the Euclidean distance of K number of neighbors. There is no standard method for determining the favorable value for K.
2. Calculate the neighbors' cases or similarities based on the distance function: the distance function has calculated the distance between each pair of observations. Euclidean is very popular and commonly used for distance measures.

$$d\,(x,y)\; = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}. \tag{30}$$

Euclidean distance in n points is considered in this formula.

We can express the above equation the following way:

$$\text{d}=\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

$d = (x_1, y_1)$ are the coordinate of one point,

(x_2, y_2) are the coordinate of another point

3. Take the K nearest neighbors as per the calculated Euclidean distance.
4. Among these K neighbors, count the number of observations that fall into each (y_i) category.
5. Classify the data point to that category into which the maximum number of the neighbors fall.

We used the caret package for KNN analysis. caret stands for Classification and Regression Training. It was developed by Max Kuhn from Pfizer in 2005. There are many important features provided in the caret package including data splitting, feature selection, feature importance, model tuning, and visualization. We normalized independent variables using Z score and max-min

methods before our analysis. The caret package provides train() for training data analysis. Predict () method is used to predict the target variable using test data.

3.19. Data Sets Constructed for Each Cancer Type

For each cancer type, four data sets were constructed for comparison purposes, each with different predictor variables…

1. Data reduction using PCs from Z-score normalized data of predictor variables
2. Data reduction using PCs from max-min normalized data of predictor variables
3. Variables selected using lasso regression
4. Variables selected using the Boruta algorithm

Data reduction using PCs from Z-score normalized data of predictor variables: First and foremost, we had 205 predictor variables after one hot encoding method. We followed the PCA technique to reduce the high dimensionality for further analysis. Before the application of the PCA technique, we normalized predictor variables using the Z score. After the PCA technique, we got 39 PCs, 30 PCS, 28 PCs, and 51 PCs predictor variables respectively for breast, lung, colon, and leukemia cancers.

Data reduction using PCs from Z-max-min normalized data of predictor variables: For the max-min normalization data set of predictor variables, we followed the same technique as the Z score. After max-min normalization and PCA technique, we gained 9 PCs for breast, 14 PCs for lung, 10 PCs for colon, and 10 PCs for leukemia cancers.

Variables selected using lasso regression: First, we prepared data and divided it into training and test data sets. We used cross-validation to determine the best lambda. After fitting a lasso regression with the scale of our dataset, we considered only those variables that have a coefficient different from zero. Discarding redundant or useless variables that have a coefficient

equal to zero. We attained 37 predictor variables for breast cancer, 33 predictors for lung cancer, 27 predictors for colon cancer, and 40 predictor variables for leukemia cancer.

Variables selected using the Boruta algorithm: Boruta algorithm is very popular for variables selection because it covered all important features including Z score, MDA, and MDG for variables selection. We used the Boruta package for analysis. We performed 99 iterations in almost 20 hours. We gained 67 predictors for breast, 52 for lung, 43 predictors for colon, and 60 predictors for leukemia cancers. How the Boruta algorithm works we already explained in the Boruta algorithm method section.

3.20. Confusion Matrix

A confusion matrix is a table that is used to summarize the performance of the binary classifier. Each row represents the predicted class, and each column represents the actual class or vice-versa. The confusion matrix is considered four basic characteristics that are used to explain the measurement of the classifier. The four characteristics are True positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Based on these characteristics, the performance metrics of an algorithm including accuracy, sensitivity, and specificity were calculated from the four characteristics. The relationship between sensitivity and specificity is illustrated through the confusion matrix or 2×2 contingency table. The following table is used in our study.

Table 5. Confusion Matrix

		Actual Condition	
Predicted Condition	Total Population	Positive	Negative
	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

3.21. Measures of Model Performance

We used three performance measures to compare the performances of our all models, including sensitivity, specificity, and accuracy. The sensitivity indicates the True Positive (TP) rate, which is the proportion of individuals who survived that were classified as surviving. The specificity refers to the True Negative (TN) rate which describes the proportion of those who did not survive that were classified as having not survived.

The formula for sensitivity, specificity, and accuracy are provided below.

$$Sensitivity = \frac{Number\ of\ True\ Positive\ (TP)}{Number\ of\ True\ Positive\ (TP) + Number\ of\ False\ Negatives\ (FN)}$$

$$Specificity = \frac{Number\ of\ True\ Negatives\ (TN)}{Number\ of\ true\ negatives\ (TN) + Number\ of\ False\ positives (FP)}$$

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

3.22. Area Under the Curve (AUC)

AUC represents a performance measurement for the classification problems at various threshold levels. It measures separability and indicates how much the model distinguishes between two classes such as survival and not survival. If the higher the AUC, the better model is at distinguishing between survival versus not survival. If AUC is close to 1, it indicates a good measure of separability. If the AUC is below 0.5, it indicates a poor model**.**

3.23. Ten (10)-Fold Cross-Validation

Cross-Validation (CV) is a resampling procedure used to evaluate machine learning techniques on a data sample to control bias and high variance. It randomly divides the set of observations into 10 groups. It is repeated with a different group of observations which is treated as a validation set. The dataset was randomly divided into 10 disjoint folds. Each fold contained

approximately the same number of records. Repeated means each fold is repeated ten times and then averaged to provide an estimate for the classifier accuracy.

The following modified formula was used in our study to estimate the CV error rate:

$$CVER = \frac{1}{10}\sum_{i=1}^{10} ERR_i, \tag{31}$$

Error rate (ERR) is computed as the number of all incorrect predictions divided by the total number of the dataset. The formula for ERR:

$$ERR_i = \frac{FP_i + FN_i}{TP_i + TN_i + FP_i + FN_i}$$

CHAPTER IV: RESULTS FOR BREAST CANCER

Table 6. Distribution of response variable for breast cancer

Categories	Frequency	Percentage
0 (Survived)	45,576	58.19
1 (Not survived)	32,744	41.81
Total	78,320	100.00

The response variable used in this analysis is a binary categorical variable with two categories: 0 and 1, where 0 indicates the patient survived and 1 indicates the patient did not survive 5 years past the initial diagnosis. The distribution of the response variable, consisting of 78,320 records, is shown in Table 6.

4.1. Data Sets for Breast Cancer Performance Measures of Different Methods

Four data sets, each with different independent variables, were used to compare the performances of the various machine learning and data reduction techniques. A comprehensive list of all potential independent variables is provided in Appendix Table 30.

Data reduction using PCs from Z-score normalized data of predictor variables: We followed the PCA technique to reduce the high dimensionality for further analysis. Before the application of the PCA technique, we normalized predictor variables using the Z score. After the PCA technique, we got 39 PCs, for breast cancer.

Data reduction using PCs from Z-max-min normalized data of predictor variables: For the max-min normalization data set of predictor variables, we followed the same technique as the Z score. After max-min normalization and PCA technique, we gained 9 PCs for breast cancer.

Variables selected using lasso regression: First, we prepared data and divided it into training and test data sets. We used cross-validation to determine the best lambda. After fitting a lasso regression with the scale of our dataset, we considered only those variables that have a

coefficient different from zero. Discarding redundant or useless variables that have a coefficient equal to zero. We attained 37 predictor variables for breast cancer.

Variables selected using the Boruta algorithm: We gained 67 predictors for breast cancer using this method.

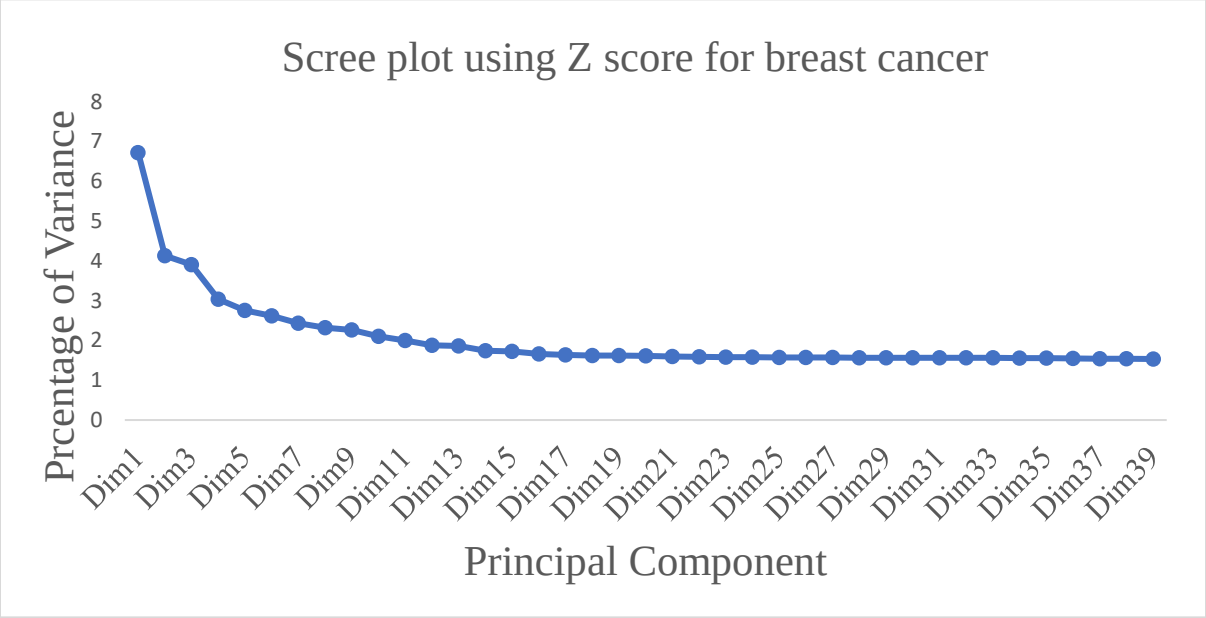

Figure 6. Percentage of variance with thirty-nine principal components using Z score for breast cancer

The first data set included principal components of the Z score normalization method, as the independent variables. For the Z score normalization method, the proportion of variance explained by the first principal component was 6.72% whereas the first two principal components explained 10.85% of the variability (Figure 6). In our study, the cumulative variance percentage was used to identify the principal component total. Furthermore, thirty-nine principal components explained 79.37% of the variability.

Table 7. Confusion matrix using PCs from the Z score normalization method in breast cancer survival prediction.

Models	Confusion Matrix		Sensitivity	Specificity
Logit model	7575	3595	0.8345	0.4542
	1502	2992		
Random forest	7295	3208	0.8036	0.5129
	1782	3379		
Decision tree	7185	3318	0.7916	0.4963
	1892	3269		
SVM linear	6921	3173	0.7625	0.5183
	2156	3414		
SVM radial	6705	3222	0.7386	0.5108
	2372	3365		
SVM sigmoid	6755	3319	0.7442	0.4961
	2322	3268		
Neural network	7330	3249	0.8075	0.5067
	1747	3338		
Naïve Bayes	7055	3290	0.7772	0.5005
	2022	3297		
LDA	7239	3479	0.7975	0.5282
	1838	3108		
KNN	7159	3396	0.7886	0.4844
	1918	3191		

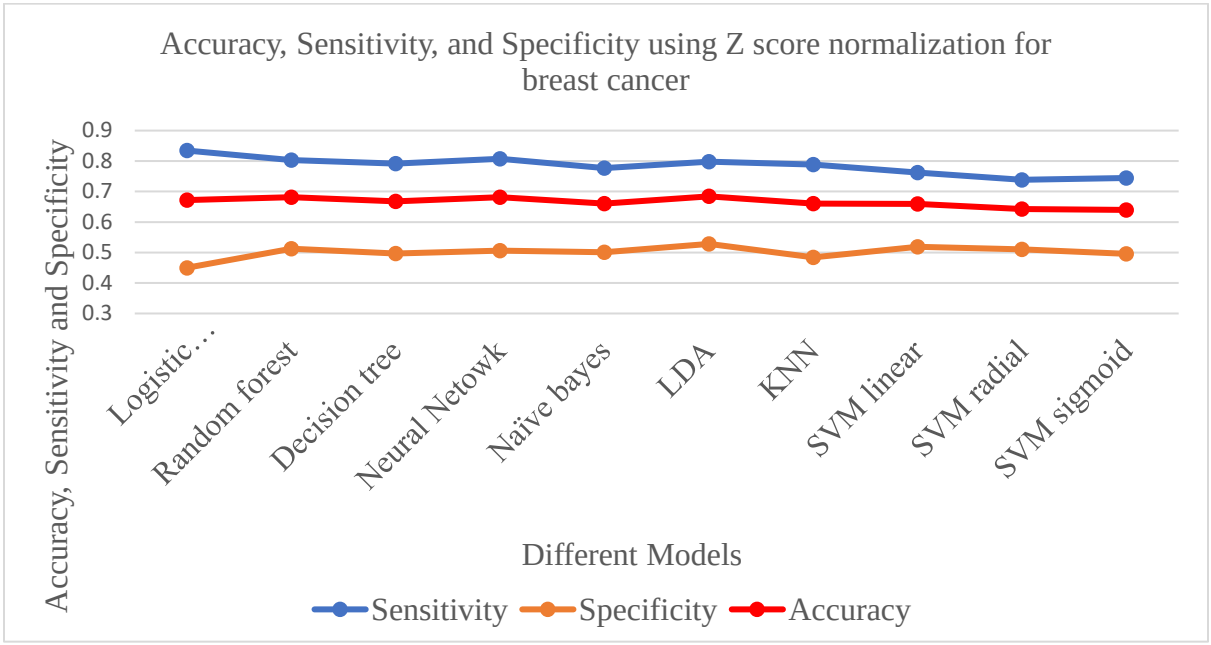

Figure 7. Comparing accuracy, sensitivity, and specificity with different methods using Z score normalization for breast cancer.

The accuracies obtained for the various machine learning techniques using the PCAs from the Z scores are illustrated in figure 7. The LDA technique had the highest accuracy of 68.42%, followed closely by the RF (68.14%) and ANN (68.10%) methods. Most of the remaining machine learning methods-maintained accuracies from 65% to 67%. The SVM sigmoid model had the lowest accuracy of 64%.

The sensitivity is the percentage of correct predictions among breast cancer patients who survived. A high sensitivity indicates a low false-negative rate. Based on the sensitivity results, LR was better than the other models with a sensitivity of 84%. The SVM radial model had the lowest sensitivity among all other models at almost 74%. The remaining methods-maintained sensitivity from 78% to 80%.

The specificity is the percentage of correct predictions among breast cancer patients who did not survive. A high specificity indicates a low false-positive rate. The LDA had the highest specificity among all other methods at almost 53%. The LR had the lowest specificity among all other methods almost at 45%. The specificity of the remaining methods ranged from 48% to 50% (Figure 7 and Table 7).

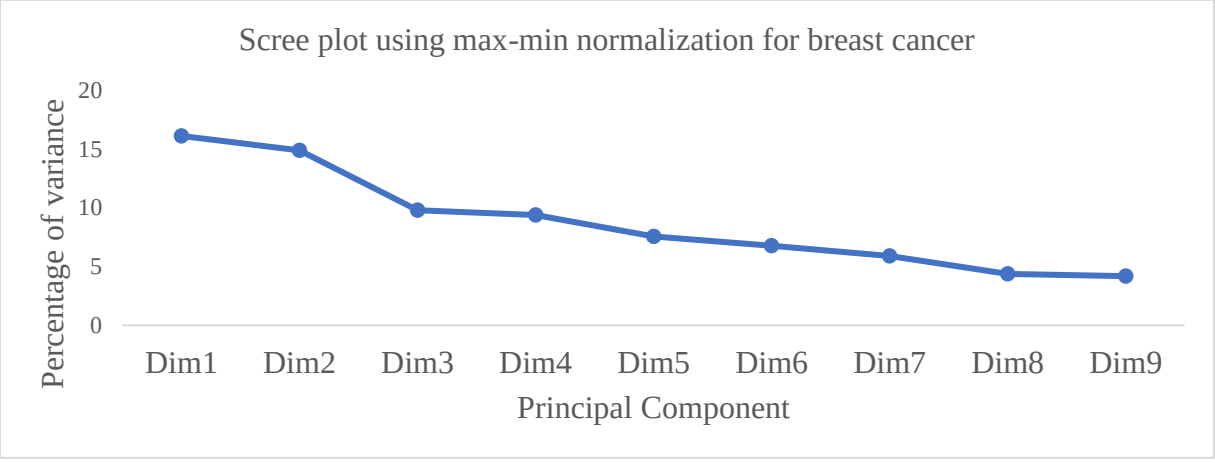

Figure 8. Percentage of variance with nine PCs using max-min normalization for breast cancer

For the Maximum-minimization normalization technique, the first principal component explained about 16.09 % of the variability whereas the first two principal components captured about 30.97% of the variability. The first nine principal components covered 78.88 % of the total variability. Therefore, we selected nine principal components for further analysis (Figure 8).

Table 8. Confusion matrix using a principal component from max-min normalization method in breast cancer survival prediction.

Models	Confusion Matrix		Sensitivity	Specificity
Logit model	7690	3472	0.8472	0.4729
	1387	3115		
Random forest	7325	3185	0.8069	0.5165
	1752	3402		
Decision tree	7205	3229	0.7937	0.5097
	1872	3358		
SVM linear	7019	3 212	0.7733	0.5124
	2058	3375		
SVM radial	6679	3321	0.7358	0.4958
	2398	3266		
SVM sigmoid	6702	3238	0.7383	0.5084
	2375	3349		
Neural network	7311	3225	0.8054	0.5103
	1766	3362		
Naïve Bayes	7113	3177	0.7836	0.5176
	1964	3410		
LDA	7183	3182	0.7913	0.5169
	1894	3405		
KNN	7067	3243	0.7785	0.5076
	2010	3344		

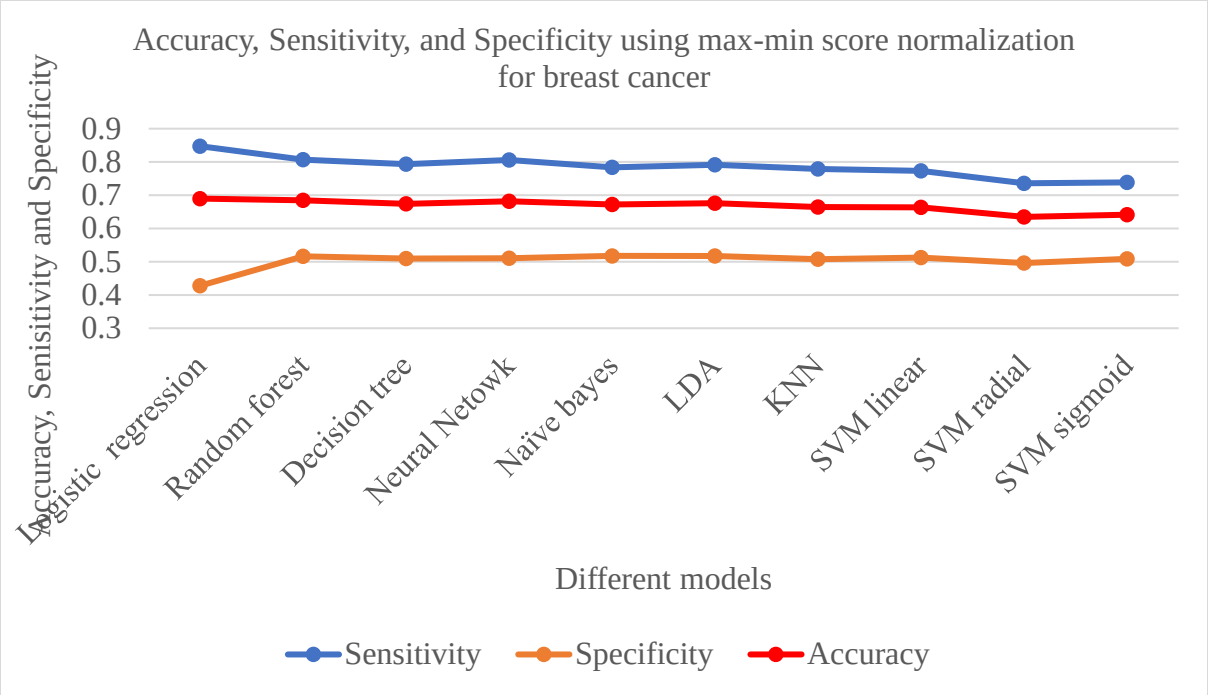

Figure 9. Compare prediction accuracy, sensitivity, and specificity with different models using max-min normalization for breast cancer.

The accuracies gained for the various machine learning techniques using the PCAs from the max-min normalization are depicted in figure 9. The LR technique had the highest accuracy of 68.97%, followed closely by the RF (68.48%) and ANN (68.14%) methods. The remaining machine learning methods-maintained accuracies from 64% to 67%. The SVM radial model had the lowest accuracy of 63.48%. LR had the highest sensitivity among all other models at almost 85%. Based on the sensitivity results, LR was better than the other models with a sensitivity of 85%. The SVM radial and sigmoid models had the lowest sensitivity among all other models at almost 74%. The sensitivity range for the remaining methods was 77% to 80%.

The NB had the highest specificity at 51.76% followed closely by the LDA (51.69%) and RF (51.65%) methods. The specificity of almost 52% illustrates that 52 % of correct predictions among breast cancer patients who did not survive. The LR had the lowest specificity among all

other methods almost at 43%. The remaining methods-maintained specificity from 49% to 50% (Figure 9 and Table 8).

Table 9. Variables selected using Boruta and Lasso regression methods for breast cancer survival prediction

Boruta	Lasso
AG, nplymnode, nlymnode, Tumor, M1, M2, M3, R1, R2, R5, R6, R7, R14, R15, R25, R27, S1, S2, S4, surg1, surg2, surg3, surg4, Radi1, Radi2, Radi3, Radi5, Radi6, g1, g2, g3, g4, H1, H2, H3, H4, H5, H6, H7, H8, H9, H10, H11, H12, H13, H14, H16, H17, H18, H19, H23, H32, prim1, prim2, prim3, prim4, prim5, prim9, Linv1, Linv2, Linv3, Linv4, Linv5, Linv6, Linv7, Linv8	AG, nplymnode, B1, B2, Tumor, M2, M3, R2, R6, R7, R10, g1, g2, S3, S4, surg2, surg5, surg6, Radi1, Radi3, Radi4, prim2, prim5, H2, H3, H4, H5, H6, H7, H8, H9, H10, H11, H12, Linv5, Linv6, Linv8

The predictor's variables description was provided in the appendix in Table 30.

Table 10. Confusion matrix using variables selection method via lasso regression in breast cancer survival prediction.

Models	Confusion Matrix		Sensitivity	Specificity
Logit model	7488	3102	0.8249	0. 5291
	1589	3485		
Random forest	7384	3012	0. 8135	0.5427
	1693	3575		
Decision tree	7129	3179	0.7854	0.5174
	1948	3408		
SVM linear	6971	2976	0.7679	0.5482
	2106	3611		
SVM radial	6995	3260	0.7706	0.5051
	2082	3327		
Sigmoid	6819	3197	0.7512	0.5147
	2258	3390		
Neural network	7402	3194	0.8155	0.5151
	1675	3393		
Naïve Bayes	7002	3087	0.7714	0.5313
	2075	3500		
LDA	7228	3079	0.7963	0.5326
	1849	3508		
KNN	6932	3139	0.7636	0.5235
	2145	3448		

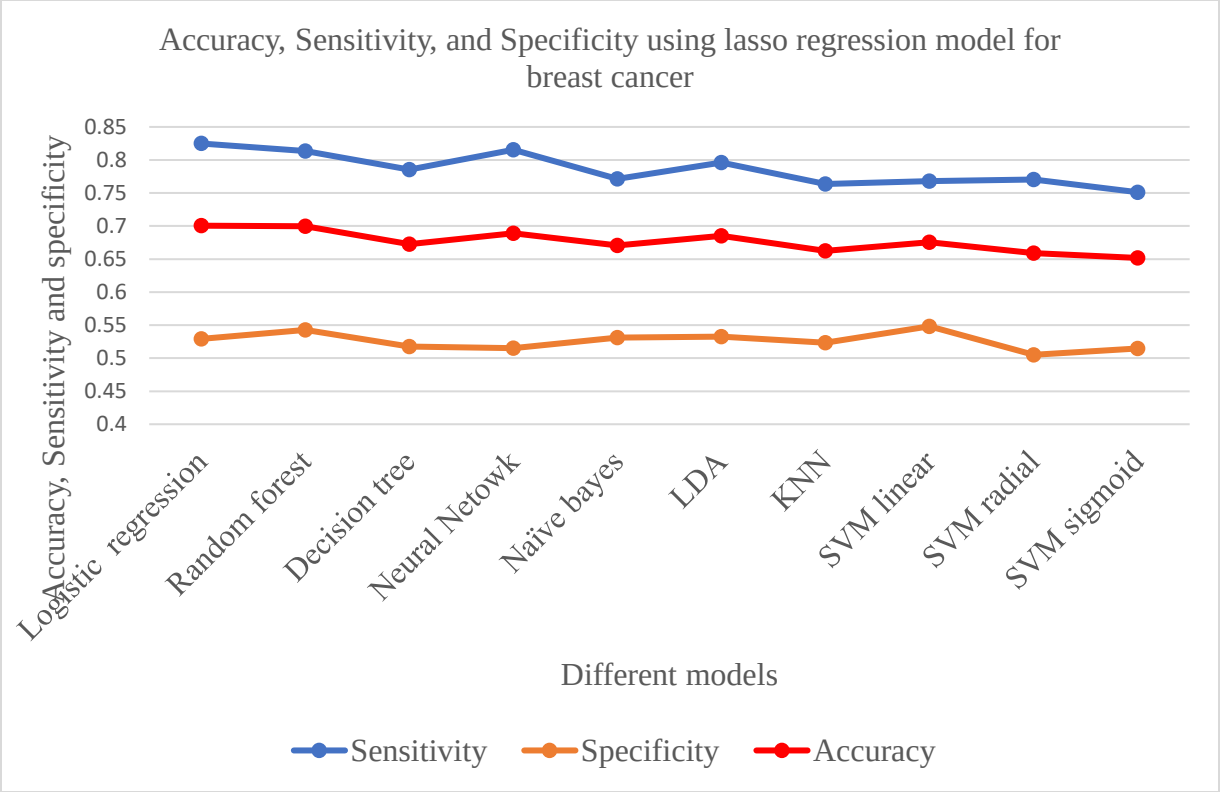

Figure 10. Comparing accuracy, sensitivity, and specificity among different models using lasso regression models for breast cancer

The comparison of accuracy was obtained for the various machine learning techniques using lasso methods. The LR and RF had reached the highest accuracy of 70%. The remaining machine learning methods-maintained accuracies from 67% to 69%. The SVM sigmoid model had the lowest accuracy of 65.17% (Figure 10).

LR had the highest sensitivity at almost 82.49% followed closely by the ANN (81.55%) and RF (81.35%). LR was better than the other models with a sensitivity of almost 83%. The SVM sigmoid model had the lowest sensitivity among all other models at almost 75.12%. The remaining methods' sensitivity ranged from 76% to 79%.

The SVM linear had the highest specificity at 54.82% followed closely by the RF (54.27%) method. The specificity of almost 55% illustrates that 55% of correct predictions among breast cancer patients who did not survive. The SVM radial had the lowest specificity among all other

methods almost at 50.51%. The specificity of other methods ranged from 51% to 53% (Figure 10 and Table 10).

Table 11. Confusion matrix using variables selected via Boruta algorithm in breast cancer survival prediction.

Models	Confusion Matrix		Sensitivity	Specificity
Logit model	7589	3011	0.8361	0.5428
	1488	3576		
Random forest	7613	3001	0.8387	0.5444
	1464	3586		
Decision tree	7601	3041	0.8374	0.5383
	1476	3546		
SVM Linear	7011	2872	0.7724	0.5639
	2066	3715		
SVM Radial	6921	2828	0.7625	0.5706
	2156	3759		
SVM Sigmoid	6881	2855	0.7581	0.5665
	2196	3732		
Neural network	7457	3123	0.8215	0.5258
	1620	3464		
Naïve Bayes	7395	3103	0.8147	0.5289
	1682	3484		
LDA	7133	3107	0.7858	0.5283
	1944	3480		
KNN	7105	3066	0.7827	0.5345
	1972	3521		

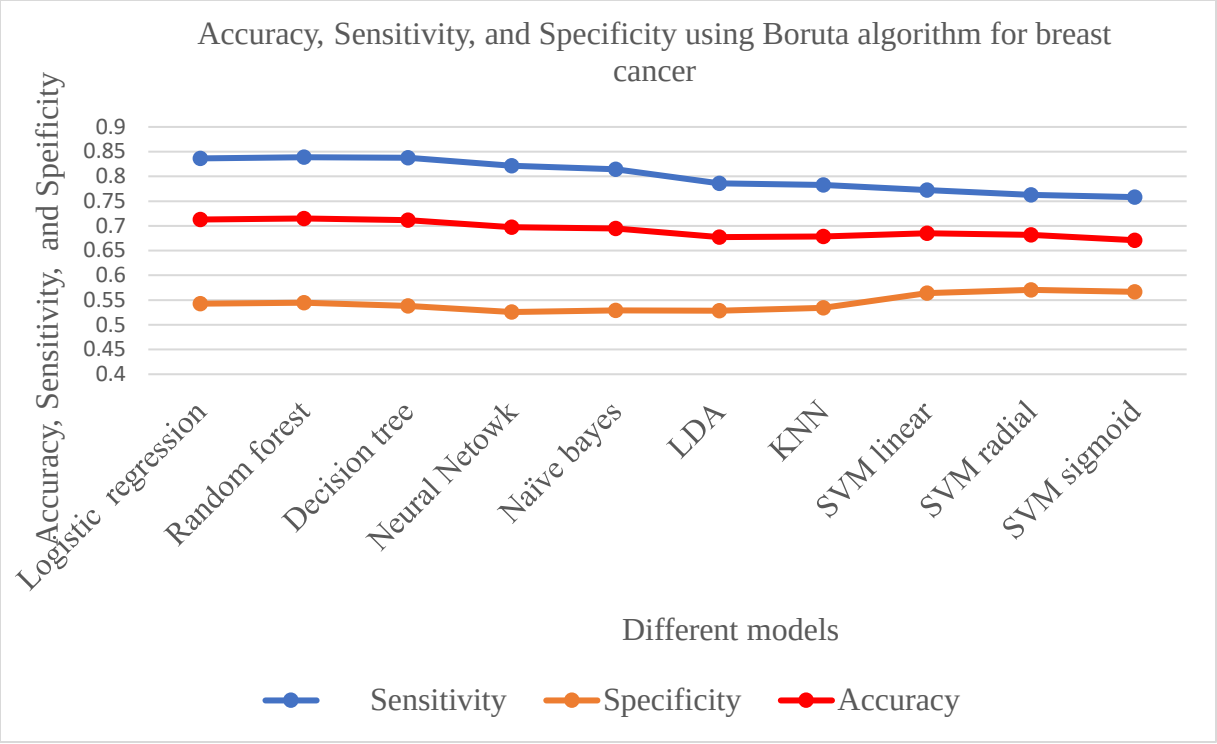

Figure 11. Comparing accuracy, sensitivity, and specificity using the Boruta algorithm for breast cancer

The comparison of accuracy obtained for the various machine learning techniques using all variables (Boruta algorithm) is presented in Figure 11. The RF had the highest accuracy of 71.49% followed closely by the LR (71.28%) and DT (71.16%) methods. The remaining machine learning methods-maintained accuracies from 67% to 68%. The LDA model had the lowest accuracy of 67.75% among all other models.

RF had the highest sensitivity at almost 83.87% followed closely by the DT (83.74%) and LR (83.61%). The sensitivity of almost 84% indicates that 84% of all the breast cancer patients that were truly survived. The SVM radial model had the lowest sensitivity among all other models at almost 75.81%. The sensitivity of other models ranged from 77% to 81%.

The SVM radial had the highest specificity at 57.06% followed closely by the SVM sigmoid (56.65%) and the SVM linear (56.39%) methods. The specificity of almost 57.06% illustrates that 57.06% of correct predictions among breast cancer patients who did not survive.

The ANN had the lowest specificity among all other methods almost at 52.58% followed closely by the LDA (52.83%) and NB (52.89%) (Figure 11 and Table 11).

4.2. Compare Area Under the Curve using Different Machine Learning Techniques and Data Reduction Techniques for Breast Cancer:

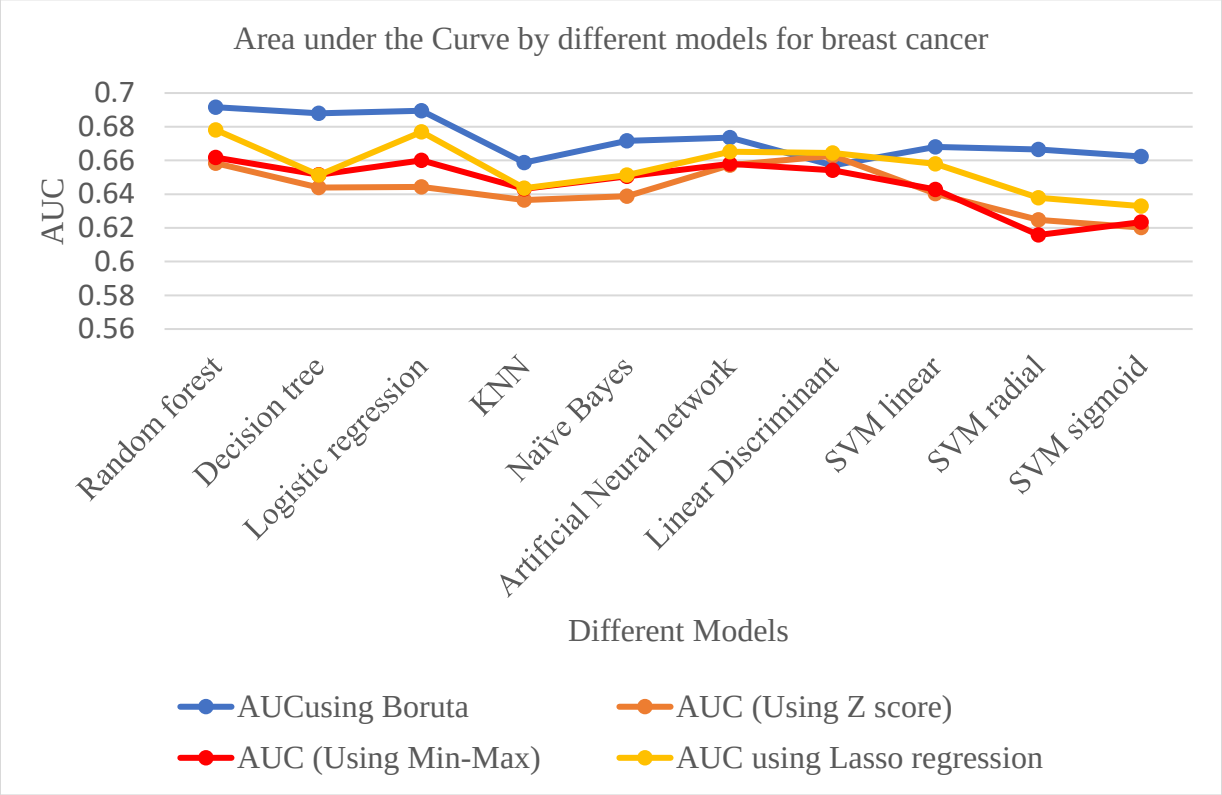

Figure 12. Compare AUC with different models using different data sets for breast cancer

The Area Under the Curve (AUC) is another way to check the classification model's performance. It tells us how much the model can differentiate between classes. The higher the AUC, the better the model is at distinguishing between patients with surviving and no survival. The RF had the highest AUC at almost 69.16% followed closely by the DT (68.79%) and LR (68.94%) using all variables (Boruta algorithm). KNN had the lowest AUC at almost 65.87%. On the other hand, LDA had the highest AUC at almost 66.28% followed closely by the RF (65.83%) and ANN (65.72%) methods using Z scores. In that case, the SVM sigmoid had the lowest AUC

at almost 62.02%. RF had the highest AUC at around 67.81% followed closely by LR (67.70%) using the lasso regression method. The SVM sigmoid had the lowest AUC at almost 63.29%. Last but not the least, RF had the highest AUC at almost 66.17% followed closely by the LR (66.01%) using the max-min normalization technique. The SVM radial had the lowest AUC at almost 61.58% (Figure 12).

CHAPTER V: RESULTS FOR LUNG CANCER

Table 12. Distribution of response variable for lung cancer

Categories	Frequency	Percentage
0 (survived)	21272.11	56.21
1 (Did not survive)	16571.89	43.79
Total	37,844	100

The response variable used in this analysis is a binary categorical variable with two categories: 0 and 1, where 0 indicates the patient survived and 1 indicates the patient did not survive 5 years past the initial diagnosis. The distribution of the response variable for lung cancer, consisting of 37,844 records, is shown in Table 12.

5.1. Data Sets for Lung Cancer Performance Measures of Different Methods

A comprehensive list of all potential independent variables is provided in appendix table 30.

Data reduction using PCs from Z-score normalized data of predictor variables: After the PCA technique, we got 30 PCS, predictor variables for lung cancer.

Data reduction using PCs from max-min normalized data of predictor variables: After max-min normalization and the PCA technique, we gained 14 PCs for lung cancer.

Variables selected using lasso regression: After fitting a lasso regression with the scale of our dataset, we attained, 33 predictors for lung cancer.

Variables selected using the Boruta algorithm: We gained 52 predictor variables for lung cancer using Boruta algorithms.

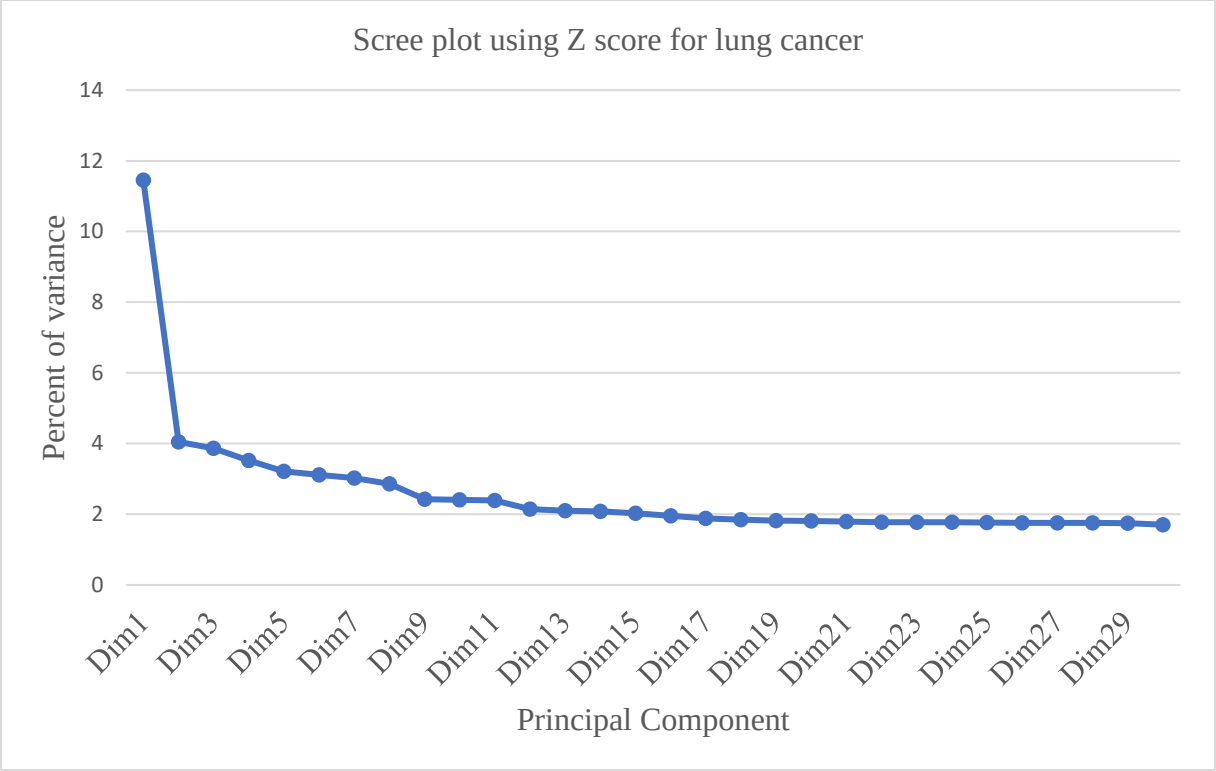

Figure 13. Percent of variance using data reduction technique PCs from Z-score normalization for lung cancer

The first data set included principal components of the Z score normalization methods; The proportion of variance explained by the first principal component is 11.45% whereas the first two principal components explained 15.50% of the variability (Figure 13). Furthermore, thirty principal components explained 79.58% of the variability.

Table 13. Confusion matrix using principal components from the Z score normalization method in lung cancer survival prediction.

Models	Confusion Matrix		Sensitivity	Specificity
Logit model	2940	1311	0.6949	0.6071
	1291	2026		
Random forest	3015	1289	0.7126	0.6137
	1216	2048		
Decision tree	2974	1298	0.7029	0.6110
	1257	2039		
SVM linear	2806	1208	0.6632	0.6379
	1425	2129		
SVM radial	2979	1314	0.7041	0.6062
	1252	2023		
SVM sigmoid	2773	1323	0.6554	0.6035
	1458	2014		
Neural network	3065	1225	0.7244	0.6329
	1166	2112		
Naïve Bayes	2806	1369	0.6632	0.5897
	1425	1968		
LDA	2902	1318	0.6859	0.6050
	1329	2019		
KNN	2822	1421	0.6669	0.5742
	1409	1916		

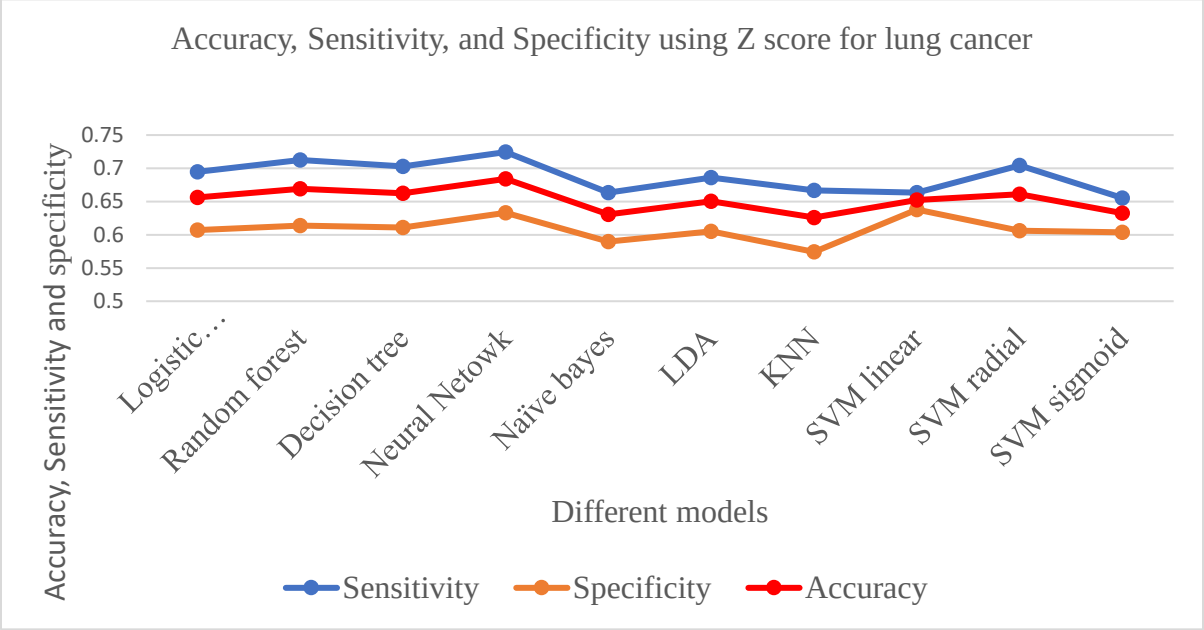

Figure 14. Comparing lung cancer accuracy, sensitivity, and specificity with different methods using Z score normalization

The accuracies gained for the various machine learning techniques using the PCAs from the Z scores are illustrated in figure 14. The ANN technique for lung cancer had the highest accuracy of 68.41%. On the other hand, the SVM sigmoid model had the lowest accuracy of 63.25%. The remaining machine learning methods covered accuracies from 66% to 67%.

The sensitivity is the percentage of correct predictions among lung cancer patients who survived. The ANN method had the highest sensitivity among all other models at almost 72.44% followed closely by the RF method (71.26%). The sensitivity of 72.44 % indicates that 72.44% of correct predictions among lung cancer patients who survived. Based on the sensitivity results, the ANN was better than other models. The SVM sigmoid model had the lowest sensitivity among all other models at almost 65.54% followed closely by SVM linear (66.32%) method. The specificity is the percentage of correct predictions among lung cancer patients who did not survive. A high specificity indicates a low false-positive rate. The SVM linear had the highest specificity among all other methods at almost 63.79% followed closely by the ANN (63.29%). It was better than other models based on the specificity. The NB method had the lowest specificity among all other methods almost at 58.97% followed closely by KNN (57.42%). The remaining methods-maintained specificity from 58% to 61% (Figure 14 and Table 13).

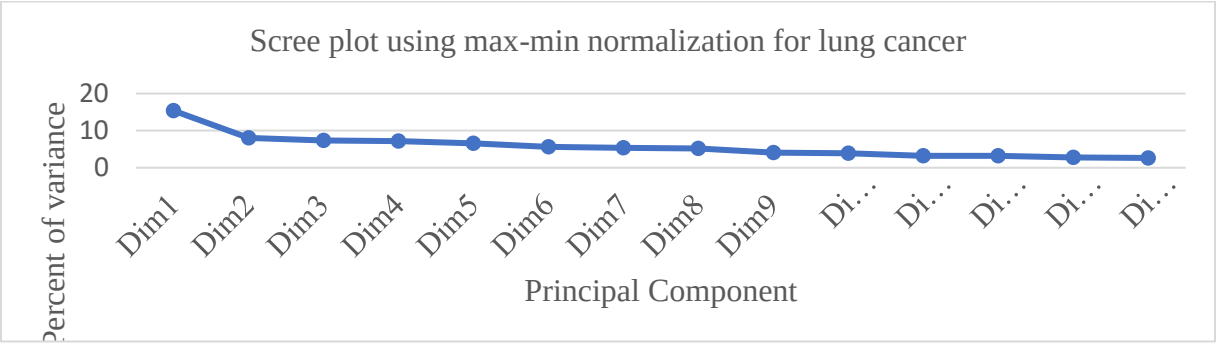

Figure 15. Percentage of variance with fourteen principal components using max-min normalization for lung cancer.

For the maximum-minimization normalization technique, the first principal component explained about 15.38% of the variability whereas the first two principal components captured about 23.44 % of the variability. The first fourteen principal components covered 80.61% of the total variability. Therefore, we selected fourteen principal components for further analysis (Figure 15).

Table 14. Confusion matrix using principal components from the max-min normalization method in lung cancer survival prediction.

Models	Confusion Matrix		Sensitivity	Specificity
Logit model	2806	1402	0.6632	0.5799
	1425	1935		
Random forest	3109	1321	0.7348	0.6041
	1122	2016		
Decision tree	2867	1307	0.6776	0.6083
	1364	2030		
SVM linear	2775	1228	0.6558	0.6320
	1456	2109		
SVM radial	2713	1239	0.6412	0.6287
	1518	2098		
SVM sigmoid	2727	1283	0.6445	0.6155
	1504	2054		
Neural network	3102	1278	0.7332	0.6170
	1129	2059		
Naïve Bayes	2792	1418	0.6599	0.5751
	1439	1919		
LDA	2932	1307	0.6929	0.6083
	1299	2030		
KNN	2805	1396	0.6639	0.5817
	1426	1941		

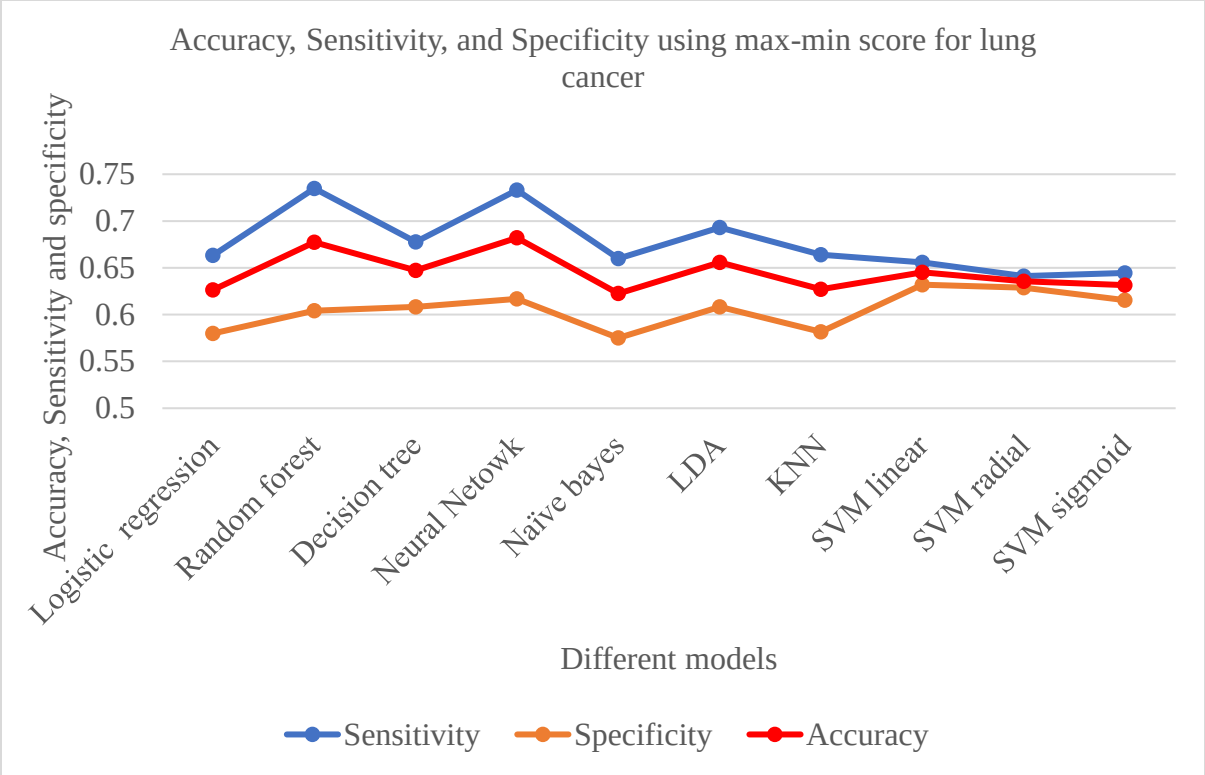

Figure 16. Comparing lung cancer accuracy, sensitivity, and specificity with different methods using max-min normalization

The RF technique for lung cancer had the highest accuracy of 67.72% using max-min normalization. On the other hand, the logit model had the lowest accuracy of 62.65% followed closely by NB (62.71%).

The RF method had the highest sensitivity among all other models at almost 73.48% followed closely by the ANN method (73.32%). The sensitivity of 73.48 % indicates that 73.48% of correct predictions among lung cancer patients who survived. Based on the sensitivity results, the RF was better than other models. The SVM radial model had the lowest sensitivity among all other models at almost 64.12% followed closely by the SVM sigmoid (64.45 %) method.

The SVM radial had the highest specificity among all other methods at almost 62.87% followed closely by the ANN (61.70%). The specificity of 62.87% indicates that 62.87% of correct predictions among lung cancer patients who did not survive. The SVM radial was better than other

models based on the specificity. The NB method had the lowest specificity among all other methods almost at 57.51% followed closely by LR (57.99 %) (Figure 16 and Table 14).

Table 15. Variables selected using Boruta and Lasso regression methods for lung cancer survival prediction

Boruta	Lasso
AG, nplymnode, nlymnode, B1, B2, Tumor, Linv1, Linv2, Linv7, Linv10, M1, M2, M5, M6, R1, R16, R24, S1, S3, S4 , S5 , E1, E15, surg1 , surg2, surg3, surg5, Radi1, Radi2, g1, g2, g3, g4, prim9, prim10, H4, H5, H17, H24, H29, H30, H37, H41, H47, H48, H52, H53, H55, H56, H62, H65, H78	AG, nplymnode, B2, Tumor, Linv5, Linv6, Linv7, M2, M5, R4, S3, S4, S5, E19, E21, R11, R16, surg1, Radi7, g1, g3, g4, H4, H18, H24, H26, H29, H47, H49, H54, H59, H78, H88,

The predictor variables were provided in the appendix in Table 30.

Table 16. Confusion matrix using the lasso method in lung cancer survival prediction.

Models	Confusion Matrix		Sensitivity	Specificity
Logit model	2811	1221	0.6644	0.6341
	1420	2116		
Random forest	2981	1161	0.7046	0.6521
	1250	2176		
Decision tree	2879	1309	0.6805	0.6077
	1352	2028		
SVM linear	2703	1386	0.6388	0.5846
	1528	1951		
SVM radial	2717	1369	0.6422	0.5896
	1514	1968		
SVM sigmoid	2749	1395	0.6497	0.5819
	1482	1942		
Neural network	2823	1206	0.6672	0.6386
	1408	2131		
Naïve Bayes	2684	1216	0.6343	0.6356
	1547	2121		
LDA	2847	1198	0.6729	0.6409
	1384	2139		
KNN	2769	1239	0.6544	0.6287
	1462			
	2098			

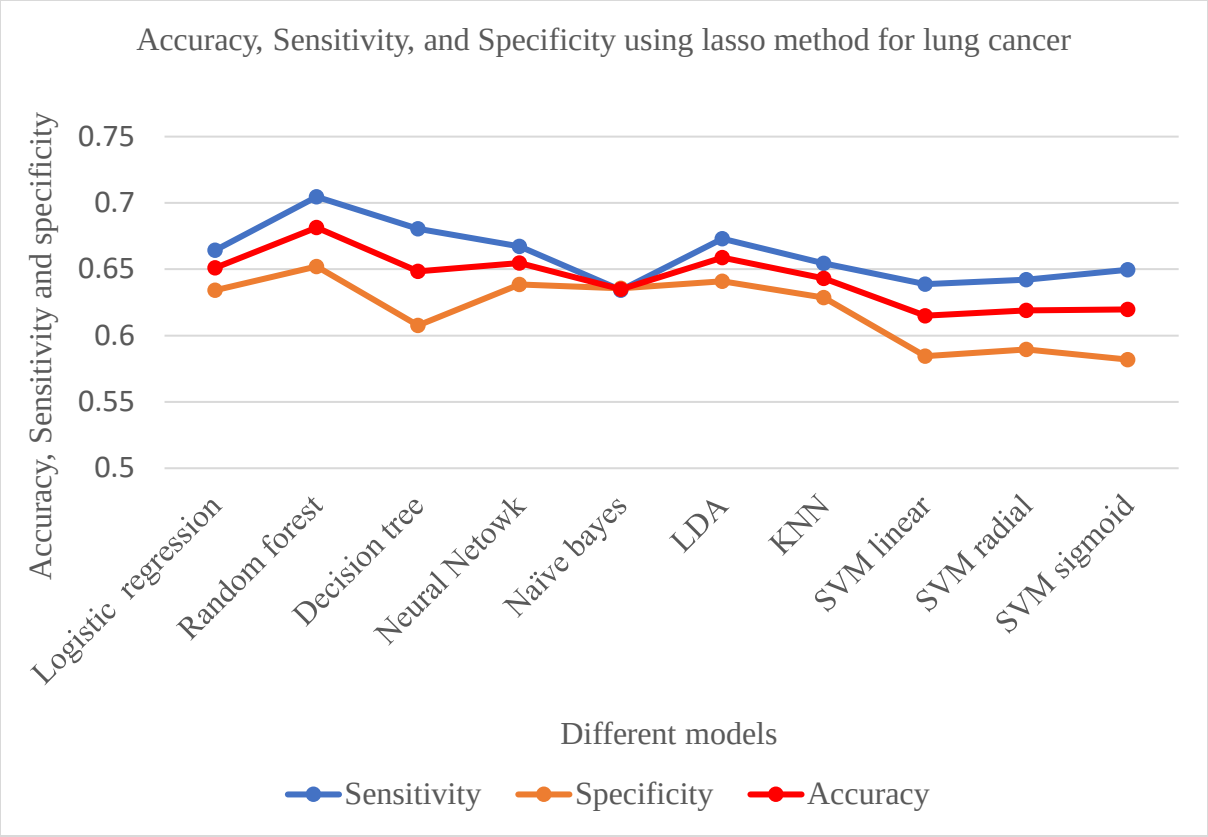

Figure 17. Comparing lung cancer accuracy, sensitivity, and specificity with different methods using the lasso method.

The RF technique for lung cancer had the highest accuracy of 68.14% using the lasso regression method. On the other hand, the LR had the lowest accuracy of 61.50% followed closely by SVM radial (61.91%), and SVM sigmoid (61.98%) methods.

The RF method had the highest sensitivity among all other models at almost 70.46%. The sensitivity of 70.46% indicates that 70.46% of correct predictions among lung cancer patients who survived. Based on the sensitivity results, the RF was better than other models. The NB model had the lowest sensitivity among all other models at almost 63.43% followed closely by SVM linear (63.88%) method.

The LDA had the highest specificity among all other methods at almost 64.09% followed closely by ANN (63.86%), and NB (63.56%) method. The specificity of 64.09% indicates that

64.09% of correct predictions among lung cancer patients who did not survive. The LDA was better than other models based on the specificity using the lasso method. The SVM sigmoid had the lowest specificity among all other methods almost at 58.19 % followed closely by SVM linear (57.46 %), and SVM radial (58.96%) methods. (Figure 17 and Table 16).

Table 17. Confusion matrix using the Boruta method in lung cancer survival prediction.

Models	Confusion Matrix		Sensitivity	Specificity
Logit model	3122	1158	0.7378	0.6529
	1109	2179		
Random forest	3205	1122	0.7575	0.6638
	1026	2215		
Decision tree	3188	1166	0.7535	0.6536
	1043	2181		
SVM linear	2912	1229	0.6883	0.6317
	1319	2108		
SVM radial	2862	1278	0.6764	0.6170
	1369	2059		
SVM sigmoid	2749	1314	0.6497	0.6062
	1482	2023		
Neural network	3203	1168	0.7570	0.6499
	1028	2169		
Naïve Bayes	2967	1206	0.7013	0.6386
	1264	2131		
LDA	3103	1215	0.7333	0.6359
	1128	2122		
KNN	3079	1198	0.7277	0.6409
	1152	2139		

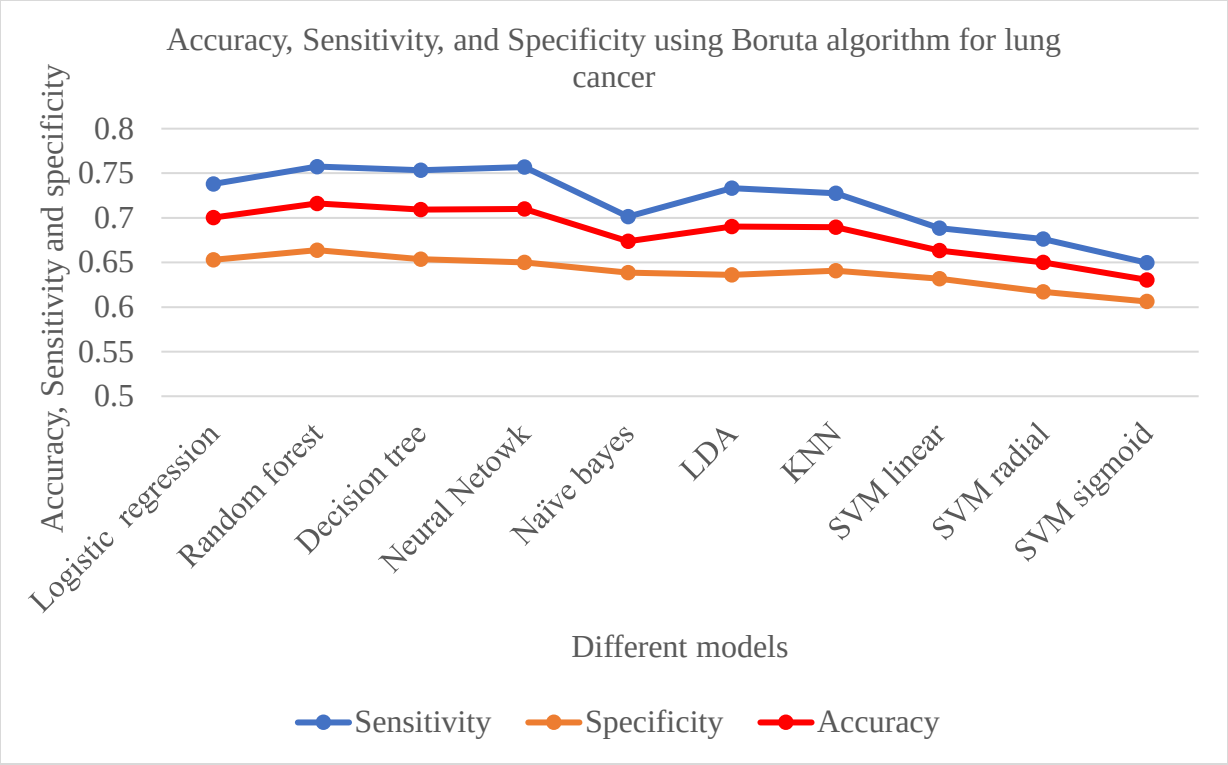

Figure 18. Comparing lung cancer accuracy, sensitivity, and specificity using the Boruta method

The accuracies gained for the various machine learning techniques using the Boruta algorithm are illustrated in figure 18. The RF technique for lung cancer had the highest accuracy of 71.62% followed closely by DT (70.94%), and LR (70.04%). On the other hand, the SVM sigmoid model had the lowest accuracy of 63.05%.

The RF method had the highest sensitivity among all other models at almost 75.75% followed closely by an ANN (75.70%), and DT (75.35%) methods. The sensitivity of 75.75% shows that 75.75% of correct predictions among lung cancer patients who survived. The SVM sigmoid model had the lowest sensitivity among all other models at almost 64.97%

The RF had the highest specificity among all other methods at almost 66.38%. The specificity of 66.38% shows that 66.38% of correct predictions among lung cancer patients who did not survive. RF was better than other models based on specificity. The SVM sigmoid method had the lowest specificity among all other methods almost at 60.62 (Figure 18 and Table 17).

5.2. Area Under the Curve for Lung Cancer Using ML and Data Reduction Methods

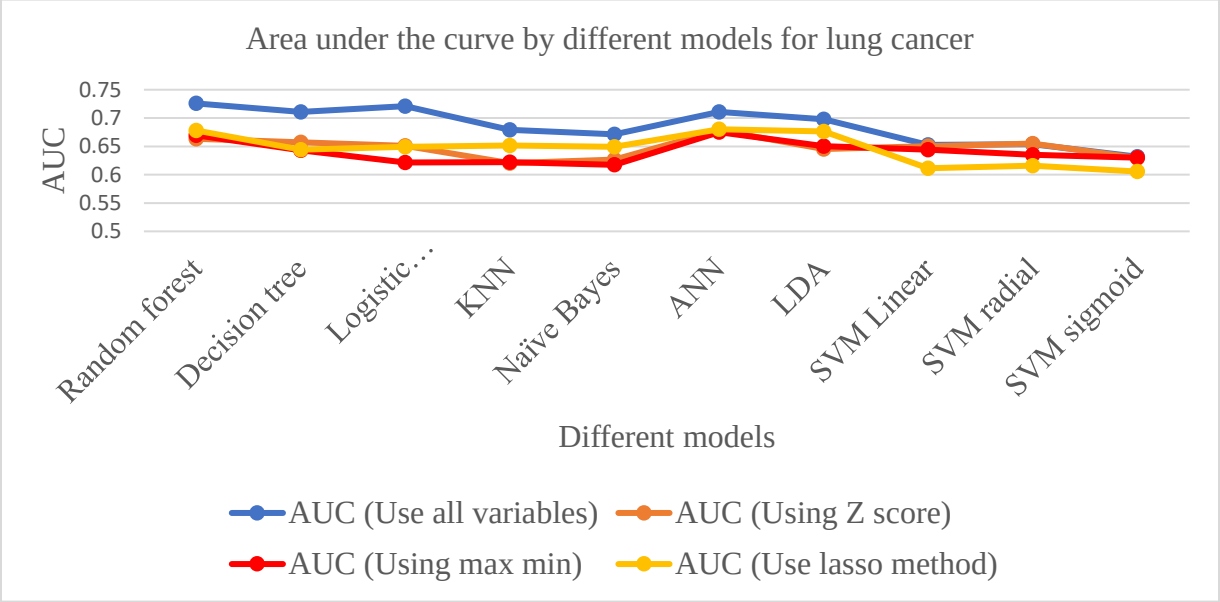

Figure 19. Compare AUC with ML and data reduction methods for lung cancer

The higher the AUC, the better the model is at distinguishing patients who survived and who did not. The RF had the highest AUC at almost 72.58% followed closely by LR (72.08%) using the variables selected method via the Boruta algorithm. The SVM sigmoid had the lowest AUC at almost 63.18%. On the other hand, ANN had the highest AUC at almost 68.02% using Z scores. In that case, the SVM sigmoid had the lowest AUC at almost 60.56%. RF had the highest AUC at around 67.83% followed closely by LDA (67.65%) using the lasso regression method. The SVM sigmoid had the lowest AUC at almost 60.56%. The ANN had the highest AUC at almost 67.50 % followed closely by random forest (66.95%) using the max-min normalization technique. (Figure 19).

CHAPTER VI: RESULTS FOR COLON CANCER

Table 18. Distribution of response variable for colon cancer

Categories	Frequency	Percentage
0 (Survived)	25,515	52.67
1 (Not survived)	22,932	47.33
Total	48,447	100.00

The response variable is a binary categorical variable with two categories: 0 and 1, where 0 indicates survived and 1 indicates not survived. The distribution of the response variable is shown in above Table 18. The five-year survived category for colon cancer consisted of 25,515 records, and 22,932 records belong to the not survival category.

6.1. Data Sets for Colon Cancer Performance Measures of Different Methods

A comprehensive list of all potential independent variables is provided in appendix table 30.

Data reduction using PCs from Z-score normalized data of predictor variables: After the PCA technique, we got 28 PCS, predictor variables for colon cancer.

Data reduction using PCs from Z-max-min normalized data of predictor variables: After max-min normalization and PCA technique, we gained, 10 PCs for colon cancer.

Variables selected using lasso regression: We attained 27 predictors for colon cancer.

Variables selected using the Boruta algorithm: We gained 43 predictors for colon cancer.

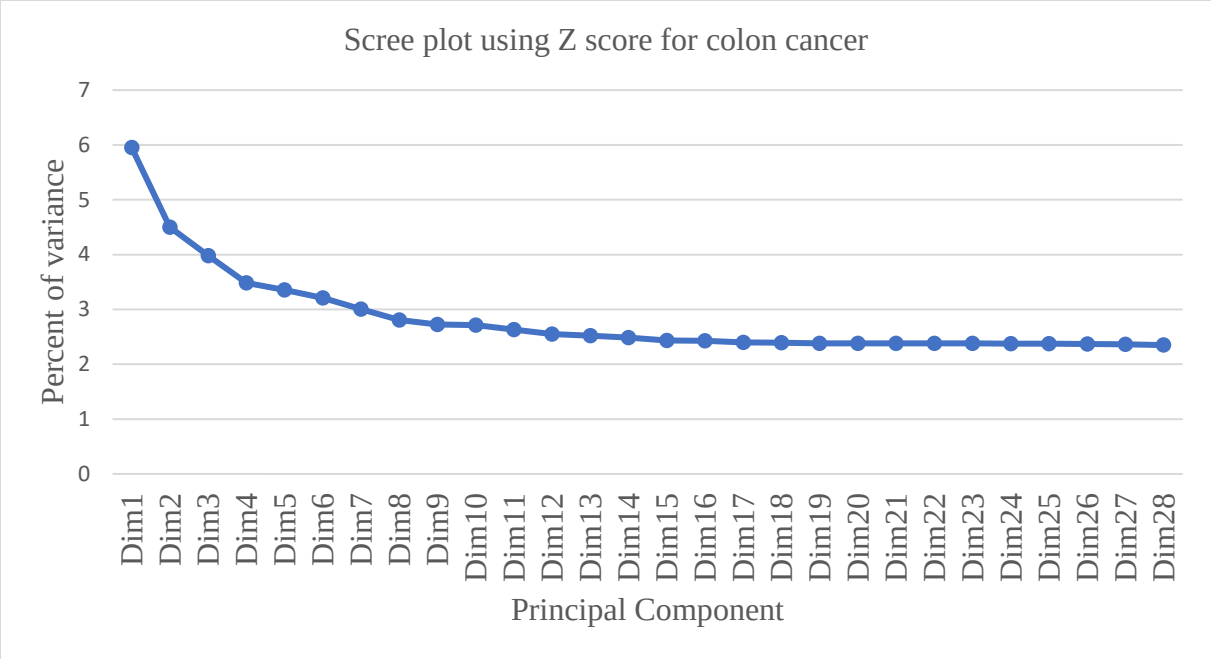

Figure 20. Percent of variance with twenty-eight principal components using Z score for colon cancer

First and foremost, the Z score is used to get the principal components. The proportion of variance explained by the first principal component is 5.95% whereas the first two principal components explained 10.45% of the variability (Figure 20). In our study, the cumulative variance percentage is used to identify the principal component total. Furthermore, twenty-eight components explained 79.32% of the variability.

Table 19. Confusion matrix using PCs from the Z score normalization method in colon cancer survival prediction.

Models	Confusion Matrix		Sensitivity	Specificity
Logit model	3208	1474	0.6346	0.6819
	1847	3160		
Random forest	3325	1696	0.6578	0.6236
	1730	2810		
Decision tree	3229	1834	0.6388	0.6042
	1826	2800		
SVM linear	2940	1899	0.5816	0.5902
	2115	2735		
SVM radial	3040	1780	0.6014	0.6158
	2015	2854		
SVM sigmoid	2880	1913	0.5697	0.5871
	2175	2721		
Neural network	3280	1730	0.6489	0.6266
	1775	2904		
Naïve Bayes	3107	1870	0.6146	0.5964
	1948	2764		
LDA	3205	1810	0.6340	0.6094
	1850	2824		
KNN	3190	1799	0.6311	0.6117
	1865	2835		

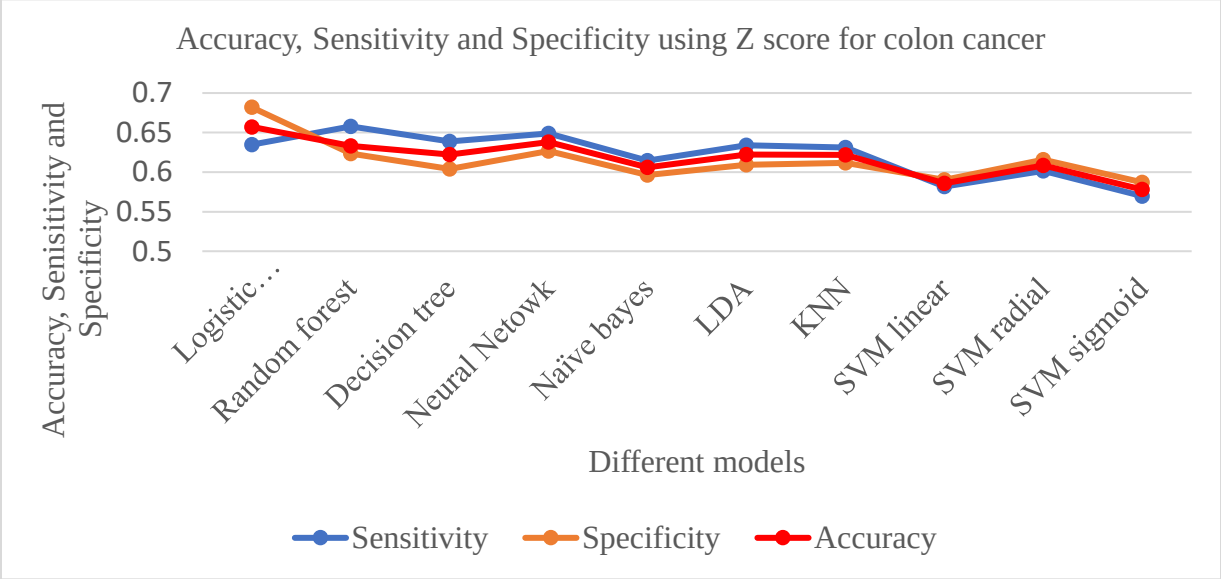

Figure 21. Comparing colon cancer accuracy, sensitivity, and specificity with different methods using Z score normalization

The accuracies gained for the various machine learning techniques using the PCAs from the Z scores are illustrated in figure 21. The logit for colon cancer had the highest accuracy of 65.72% On the other hand, the SVM sigmoid model had the lowest accuracy of 57.80% followed closely by the SVM linear (58.70%).

The RF method had the highest sensitivity among all other models at almost 65.78% followed closely by the ANN method (64.89%). The sensitivity of 65.78% indicates that 65.78% of correct predictions among colon cancer patients who survived. The SVM sigmoid model had the lowest sensitivity among all other models at almost 56.97%

The LR had the highest specificity among all other methods at almost 68.19%. The specificity of 68.19% indicates that 68.19% of correct predictions among colon cancer patients who did not survive. The SVM sigmoid method had the lowest specificity among all other methods almost at 58.71% followed closely by SVM linear (59.02%). Based on the specificity, LR had the best method using the z score (Figure 21 and Table 19).

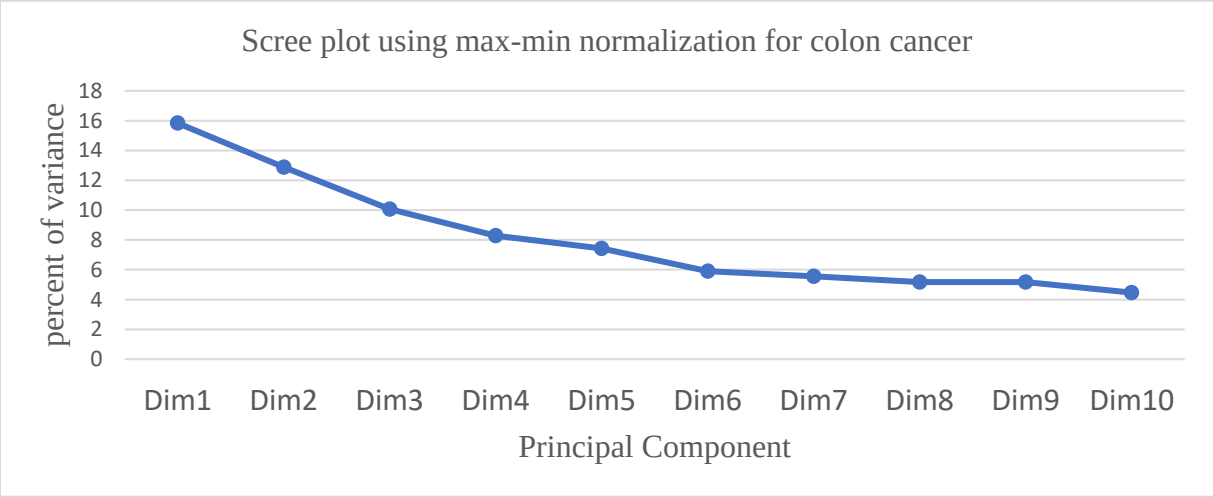

Figure 22. Percentage of variance with ten principal components using max-min normalization for colon cancer

The first principal component explained about 15.83% of the variability whereas the first two principal components explained 28.71% of the variability (Figure 22). Moreover, the first ten

principal components covered 80.77 % of the total variability. Therefore, we selected ten principal components for further analysis.

Table 20. Confusion matrix using PCs from the max-min normalization method in colon cancer survival prediction.

Models	Confusion Matrix		Sensitivity	Specificity
Logit model	2965	1462	0.5865	0.6845
	2090	3172		
Random forest	3190	1886	0.6311	0.5930
	1865	2748		
Decision tree	3164	1821	0.6259	0.6070
	1891	2813		
Linear	2907	1842	0.5750	0.6025
	2148	2792		
Radial	3070	1765	0.6073	0.6191
	1985	2869		
Sigmoid	3013	1777	0.5960	0.6165
	2042	2857		
Neural network	3170	1896	0.6271	0.5908
	1885	2738		
Naïve Bayes	3020	1905	0.5974	0.5889
	2035	2729		
LDA	3085	1970	0.6102	0.5748
	1970	2664		
KNN	3075	1927	0.6083	0.5842
	1980	2707		

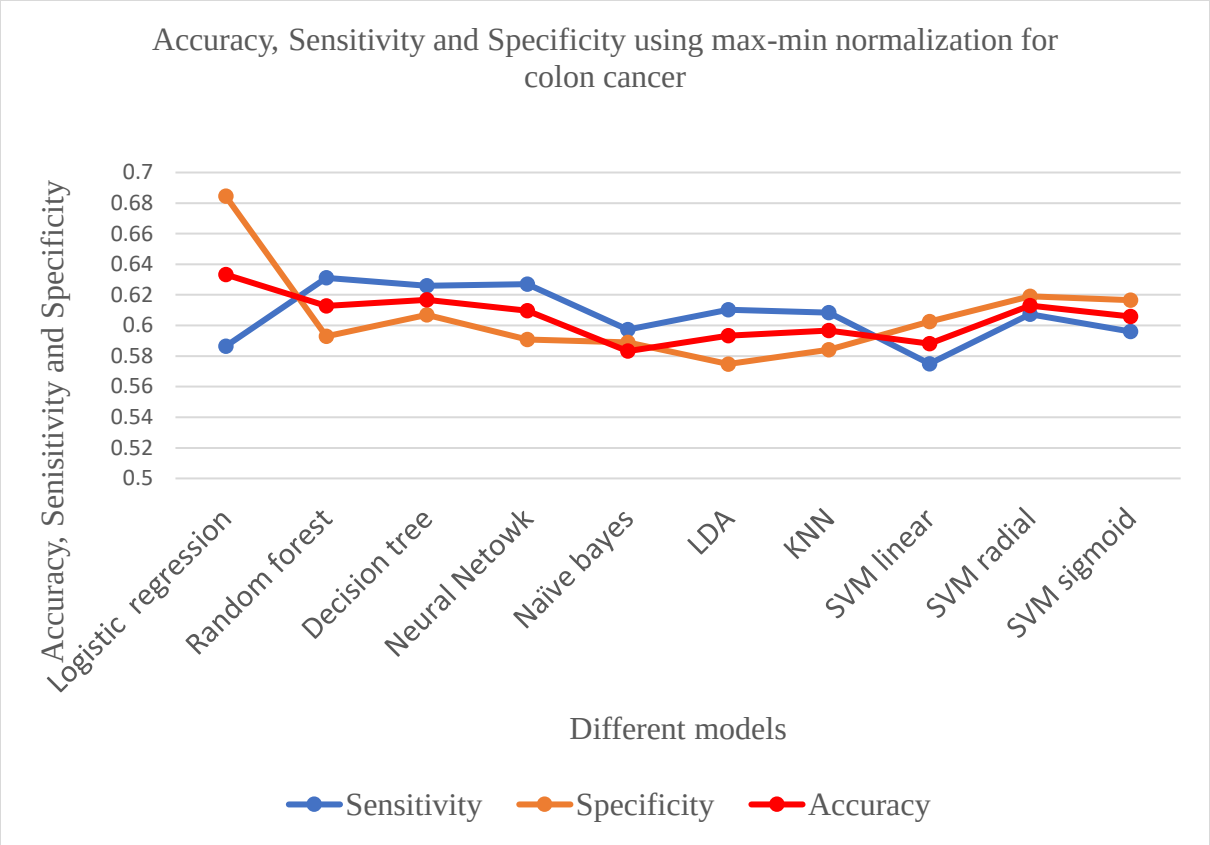

Figure 23. Comparing colon cancer accuracy, sensitivity, and specificity with different methods using max-min normalization

The accuracies gained for the various machine learning techniques using the PCAs from the max-min are illustrated in figure 23. The LR for colon cancer had the highest accuracy of 63.33%. On the other hand, the SVM linear model had the lowest accuracy at 58.82%.

The RF method had the highest sensitivity among all other models at almost 63.11% followed closely by the DT method (62.59%). The sensitivity of 63.11% indicates that 63.11% of correct predictions among colon cancer patients who survived. The SVM linear model had the lowest sensitivity among all other models at almost 57.50% followed closely by SVM sigmoid (59.60%), and NB (59.74%) respectively.

The LR had the highest specificity among all other methods at almost 68.45%. The specificity of 68.45% indicates that 68.45% of correct predictions among colon cancer patients

who did not survive. The LDA had the lowest specificity among all other methods almost at 57.48%. Based on the specificity, LR had the best method using max min (Figure 23 and Table 20).

Table 21. Variables selection using Boruta and Lasso regression methods

Boruta	Lasso
AG,B1,Tumor,M1,M2,M4,R1,R2,R4,R5,R6,S1,S2,S3,Radi6, Radi8, surg1,surg2,g1,g2,g3,H1,H2,H3,H4,H5,H6,H7, H68, H76,H80,E12,E15,E24,Linv4,Linv5,Linv6,Linv7,Linv8,Linv9, prim1, prim2,prim3	AG, Tumor, M2, M5, R2, g1, g5, S3 , S4, surg5 , Radi1, Radi3, Radi4, H2, H3, H4, H45, H62, H77, H83, Linv5, Linv6, Linv8, prim2,prim8, prim9, E18

The predictor variables were provided in the appendix in Table 30.

Table 22. Confusion matrix using variable selection method via lasso regression for colon cancer.

Models	Confusion Matrix		Sensitivity	Specificity
Logit model	3496	1746	0.6915	0.6232
	1559	2888		
Random forest	3529	1736	0.6981	0.6253
	1526	2898		
Decision tree	3455	1786	0.6834	0.6145
	1600	2848		
SVM linear	3153	1807	0.6237	0.6101
	1902	2827		
SVM radial	3109	1796	0.6150	0.6124
	1946	2838		
SVM sigmoid	3123	1868	0.6178	0.5969
	1932	2766		
Neural network	3514	1708	0.6951	0.6314
	1541	2926		
Naïve Bayes	3262	1748	0.6453	0.6227
	1793	2886		
LDA	3317	1733	0.6562	0.6260
	1738	2901		
KNN	3311	1728	0.6549	0.6271
	1744	2906		

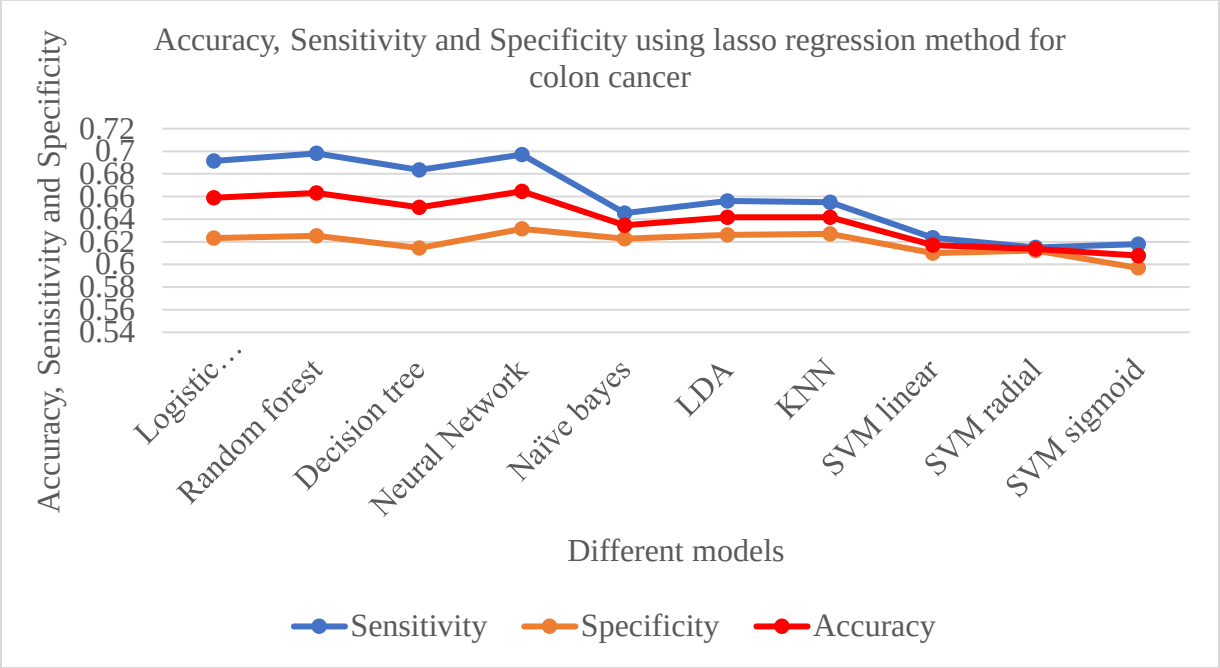

Figure 24. Comparing colon cancer accuracy, sensitivity, and specificity with different methods using the lasso regression method

The accuracies attained for the various machine learning techniques using the PCAs from the lasso regression method. The ANN for colon cancer had the highest accuracy of 66.46% followed closely by RF (66.33%), and LR (65.88%) respectively. On the other hand, the SVM sigmoid model had the lowest accuracy of 60.78%.

The RF method had the highest sensitivity among all other models at almost 69.81% followed closely by the ANN (69.51%), and LR (69.15%). The sensitivity of 69.81 % indicates that 69.81% of correct predictions among colon cancer patients who survived. The SVM radial model had the lowest sensitivity among all other models at almost 61.50% followed closely by the SVM sigmoid (61.78%).

The ANN had the highest specificity among all other methods at almost 63.14% followed closely by LDA (62.60%), KNN (62.71%), and RF (62.53%) respectively. The specificity of 63.14% indicates that 63.14% of correct predictions among colon cancer patients who did not

survive. The SVM sigmoid had the lowest specificity among all other methods almost at 59.69% (Figure 24 and Table 22).

Table 23. Confusion matrix using selected variables via Boruta algorithm for colon cancer.

Models	Confusion Matrix		Sensitivity	Specificity
Logit model	3619	1628	0.7159	0.6486
	1436	3006		
Random forest	3835	1730	0.7586	0.6266
	1220	2904		
Decision tree	3539	1860	0.7000	0.5986
	1516	2774		
SVM linear	3353	1514	0.6633	0.6732
	1702	3120		
SVM radial	3826	1766	0.7568	0.6189
	1229	2868		
SVM sigmoid	3849	2114	0.7614	0.5438
	1206	2520		
Neural network	3985	1650	0.7883	0.6439
	1070	2984		
Naïve Bayes	3709	1756	0.7337	0.6211
	1346	2878		
LDA	3792	1702	0.7501	0.6327
	1263	2932		
KNN	3765	1855	0.7448	0.5996
	1290	2779		

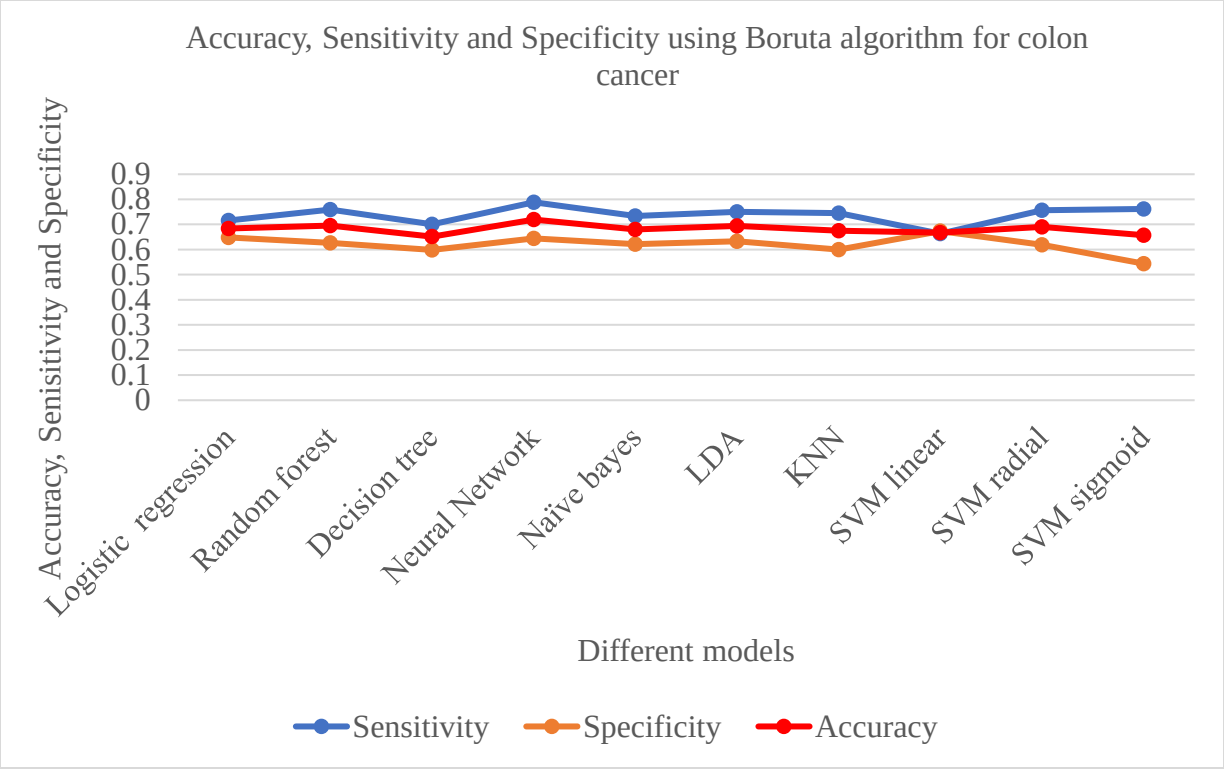

Figure 25. Comparing colon cancer accuracy, sensitivity, and specificity using the Boruta algorithm

The accuracies were obtained for the various machine learning techniques using the Boruta algorithm. The ANN for colon cancer had the highest accuracy of 71.93%. On the other hand, the DT model had the lowest accuracy of 65.15%.

The ANN method had the highest sensitivity among all other models at almost 78.83%. The sensitivity of 78.83% indicates that 78.83% of correct predictions among colon cancer patients who survived. The SVM linear model had the lowest sensitivity among all other models at almost 66.33%. The ANN was the best model based on sensitivity.

The SVM linear had the highest specificity among all other methods at almost 67.32%. The specificity of 67.32% indicates that 67.32% of correct predictions among colon cancer patients who did not survive. The SVM sigmoid had the lowest specificity among all other methods almost at 54.38 % (Figure 25 and Table 23).

6.2. Area Under the Curve for Colon Cancer Using ML and Data Reduction Methods

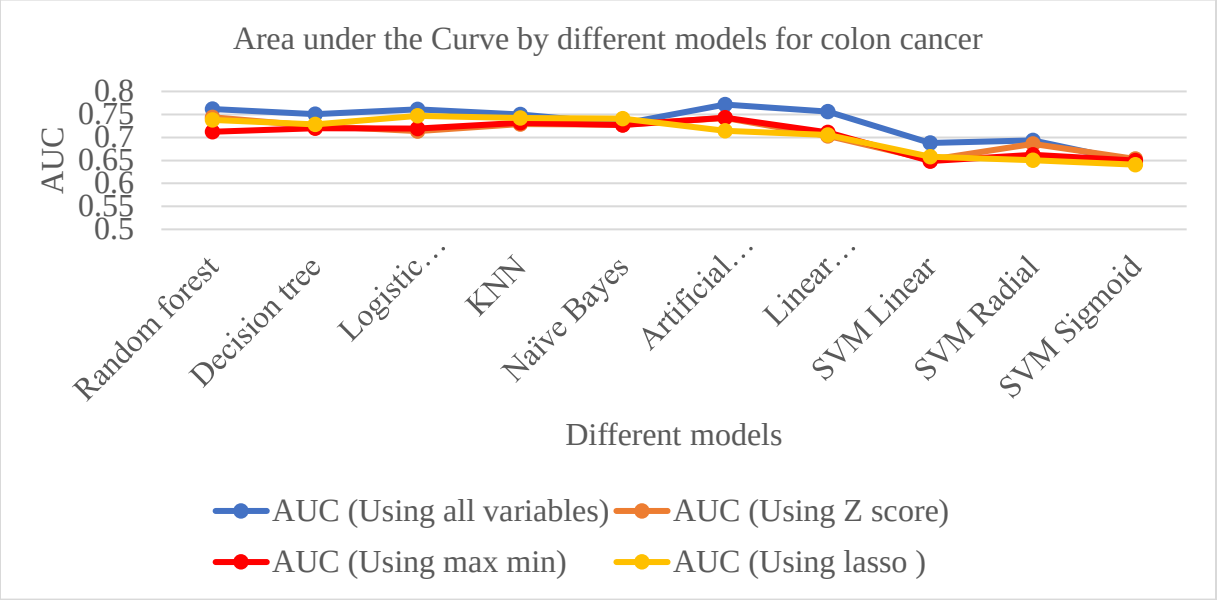

Figure 26. Compare AUC for colon cancer with different models using different variables sets

The AUC was obtained for the various machine learning techniques using four different datasets with different independent variables. The ANN had the highest AUC at almost 77.15% followed closely by LR (76.11%), and RF (76.16%) using all variables selected technique via the Boruta algorithm. On the other hand, the RF had the highest AUC at almost 74.38% followed closely by ANN (74.24%) using Z scores. LR had the highest AUC at around 74.68 % using the lasso regression method. The ANN had the highest AUC at almost 74.29% using the data reduction method from the max-min normalization technique. (Figure 26).

CHAPTER VII: RESULTS FOR LEUKEMIA CANCER

Table 24. Distribution of response variable for leukemia cancer

Categories	Frequency	Percentage
0 (Survived)	55,956	58.15
1 (Not survived)	40,271	41.85
Total	96,227	100.00

The response variable is a binary categorical variable with two categories: 0 and 1, where 0 indicates survived and 1 indicates not survived. The distribution of the response variable is shown in table 24. The five-year survived category for leukemia cancer consisted of 55,956 records, and 40,271 records belong to the not survival category.

7.1. Data Sets for Leukemia Cancer Performance Measures of Different Methods

Data reduction using PCs from Z-score normalized data of predictor variables: After the PCA technique, we got 51 PCs predictor variables for leukemia cancer.

Data reduction using PCs from Z-max-min normalized data of predictor variables: After max-min normalization and PCA technique, we gained 10 PCs for leukemia cancer.

Variables selected using lasso regression: We attained 40 predictor variables for leukemia cancer. Variables selected using the Boruta algorithm: We gained 60 predictors for leukemia cancer.

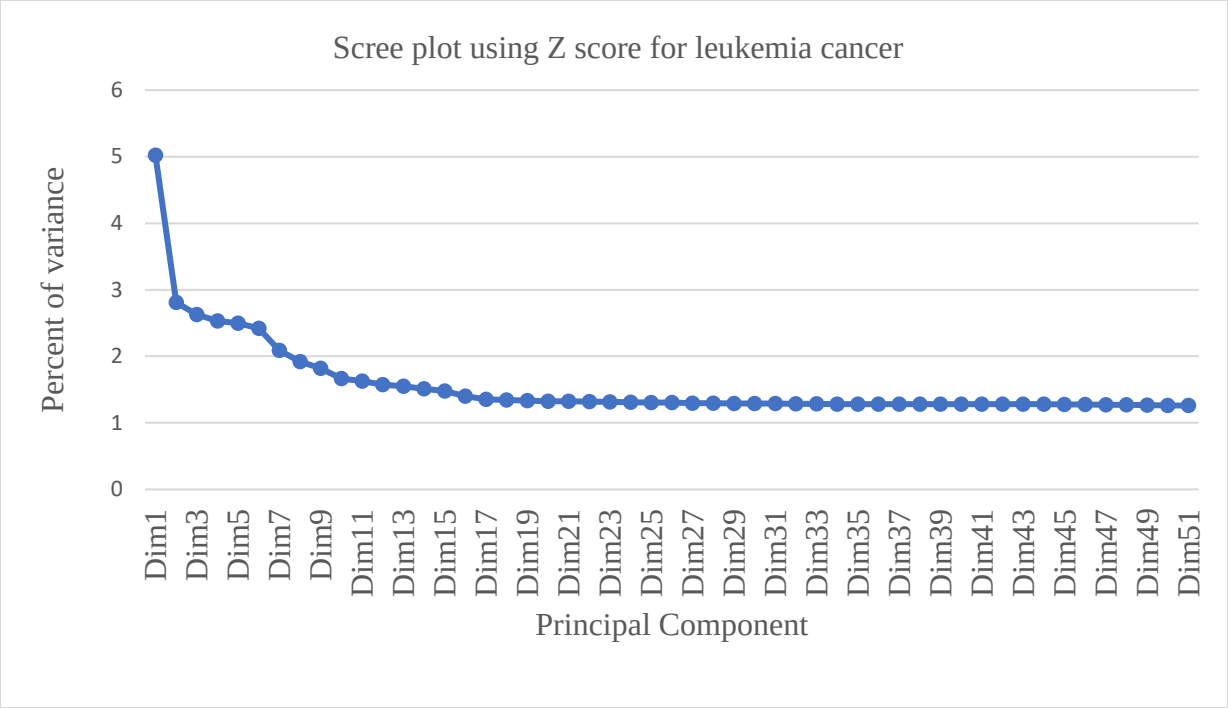

Figure 27. Percentage of variance with fifty-one principal components using Z score for leukemia cancer

The first data set included principal components of the Z score normalization method. the proportion of variance explained by the first principal component was 5.02% whereas the first two principal components explained 7.83% of the variability (Figure 27). Furthermore, fifty-one principal components explained 80.16% of the variability.

Table 25. Confusion matrix using PCs from the Z score normalization method in leukemia cancer survival prediction

Models	Confusion Matrix		Sensitivity	Specificity
Logit model	9870	2355	0.8855	0.7092
	1275	5745		
Random forest	9906	2275	0.8888	0.7191
	1239	5825		
Decision tree	9689	2244	0.8693	0.7229
	1456	5856		
Linear	8655	2472	0.7765	0.6948
	2490	5628		
Radial	8632	2426	0.7745	0.7004
	2513	5674		
Sigmoid	8721	2396	0.7825	0.7042
	2424	5704		
Neural network	9802	2175	0.8794	0.7314
	1343	5925		
Naïve Bayes	8954	2297	0.8034	0.7164
	2191	5803		
LDA	9367	2235	0.8404	0.7240
	1778	5865		
KNN	9261	2211	0.8309	0.7270
	1884	5889		

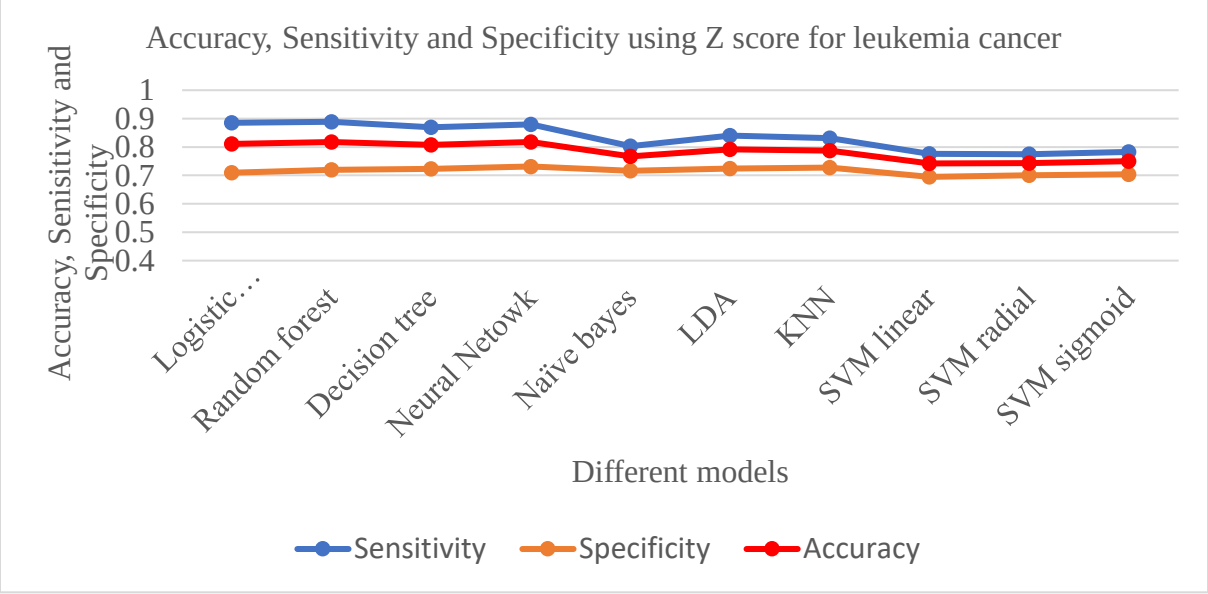

Figure 28. Comparing leukemia cancer accuracy, sensitivity, and specificity with different methods using Z score normalization

The accuracies obtained for the various machine learning techniques using the PCAs from the Z scores are illustrated in figure 27. The ANN had the highest accuracy of 81.72% followed closely by the RF (81.74%) and LR (81.13%) respectively. The SVM linear model had the lowest accuracy of 74.21% followed closely by the SVM radial (74.33%) (Figure 28).

The sensitivity is the percentage of correct predictions among breast cancer patients who survived. Based on the sensitivity results, the RF was better than the other models with a sensitivity of 88.88% followed closely by LR (88.55%). The SVM radial model had the lowest sensitivity among all other models at almost 77.45% followed closely by SVM linear (77.65%).

The specificity is the percentage of correct predictions among breast cancer patients who did not survive. A high specificity indicates a low false-positive rate. The ANN had the highest specificity among all other methods at almost 73.14% followed closely by KNN (72.70%). The SVM linear had the lowest specificity among all other methods almost at 69.48% (Figure 28 and Table 25).

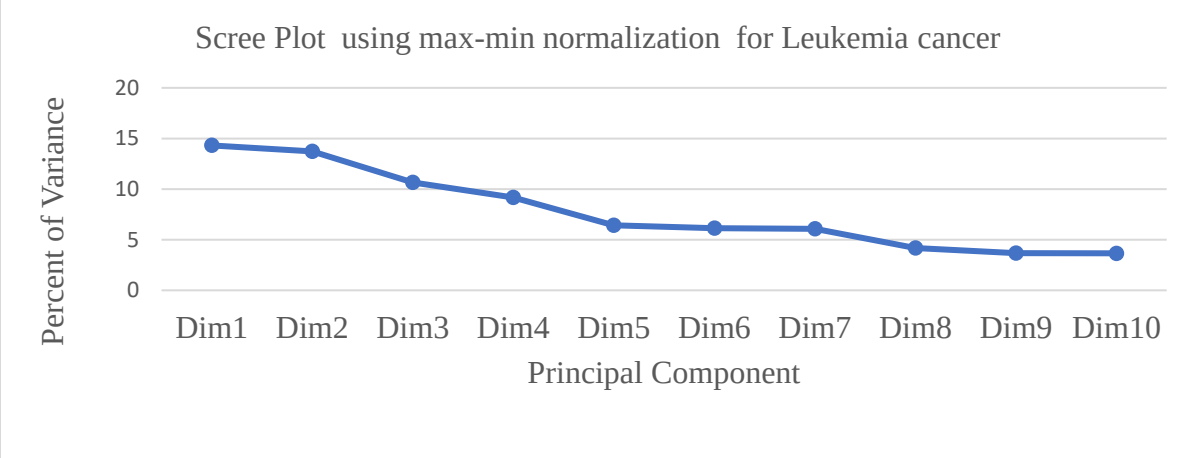

Figure 29. Percentage of variance with ten principal components using max-min normalization for leukemia cancer

For the maximum-minimization normalization technique, the first principal component explained about 14.31% of the variability whereas the first two principal components captured

about 28.04 % of the variability. The first fourteen principal components covered 78.03% of the total variability. Therefore, we have selected ten principal components for further analysis (Figure 29).

Table 26. Confusion matrix using PCs from the max-min normalization method in leukemia cancer survival prediction.

Models	Confusion Matrix		Sensitivity	Specificity
Logit model	9799	2208	0.8792	0.7274
	1346	5892		
Random forest	9876	2183	0.8861	0.7305
	1269	5917		
Decision tree	9814	2193	0.8805	0.7292
	1331	5907		
SVM linear	8574	2482	0.8757	0.6935
	2571	5618		
SVM radial	8613	2354	0.7728	0.7093
	2532	5746		
SVM sigmoid	8529	2396	0.8633	0.7042
	2616	5704		
Neural network	9896	2134	0.8879	0.7365
	1249	5966		
Naïve Bayes	9776	2332	0.8771	0.7121
	1369	5768		
LDA	9821	2331	0.8812	0.7122
	1324	5769		
KNN	9771	2366	0.8767	0.7079
	1374	5734		

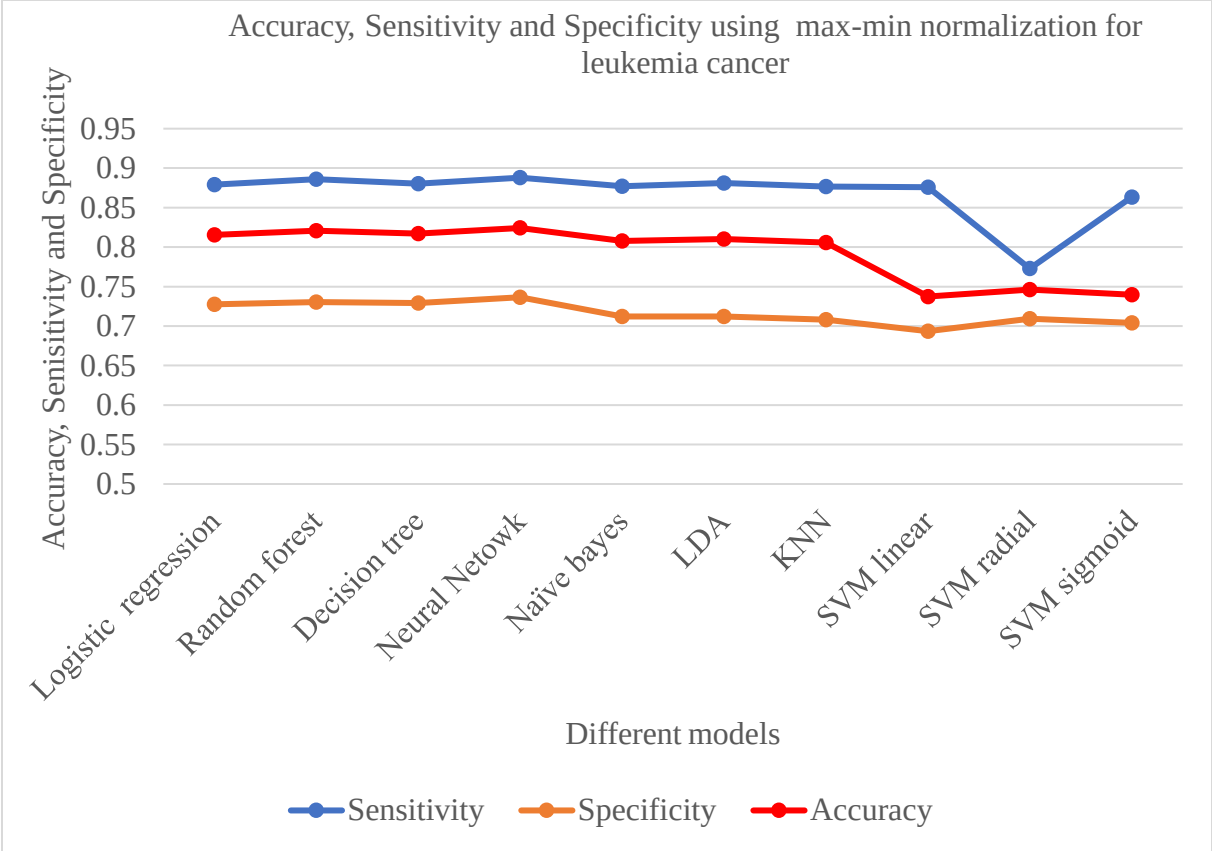

Figure 30. Comparing leukemia cancer accuracy, sensitivity, and specificity with different methods using max-min normalization

The accuracies obtained for the various machine learning techniques using the PCAs from the max-min normalization are illustrated in figure 28. The ANN had the highest accuracy of 82.42% followed closely by the RF (82.06%) and LR (81.53%) respectively. The SVM sigmoid model had the lowest accuracy of 73.96% followed closely by SVM linear (73.74%) (Figure 30).

ANN was better than the other models with a sensitivity of 88.79% followed closely by RF (88.61%). The SVM radial model had the lowest sensitivity among all other models at almost 77.28%.

The ANN had the highest specificity among all other methods at almost 73.65% followed closely by RF (73.05%). The KNN had the lowest specificity among all other methods almost at 57.42% (Figure 30 and Table 26).

Table 27. Variables selection using Boruta and Lasso regression methods for leukemia cancer

Boruta	Lasso
AG , nplymnode , B1 ,Tumor,M1,M2, M3, M4, E1, E4, E7,E9, E14,E21,S4,surg1,surg2,surg3,Radi1,Radi2,Radi3,Radi4,Radi5, g1,g2,g3,H1,H2,H3,H4,H5,H6,H7,H8,H9,H10,H11, H12,H13,H14,H15,H26,H37,H48,H49,H56,H57,H62 ,H65,Linv3,Linv4,Linv5,Linv6,Linv7,Linv8, Linv9, prim1, prim2,prim3,prim4	AG, M2, M3, R2, R3,R4,R5,R6,R27, g1, g2, g5, S3 , S4, S5, E19, E20, E24, surg2 , Radi3, Radi4, Radi7,Radi9, H2, H3, H4, H5,H6,H7,H8,H36,H47,H68,H69,Linv2,Linv3,Linv5,Linv6, prim6,prim8

The predictor variables were provided in the appendix in Table 30.

Table 28. Confusion matrix using variable selection by lasso regression for leukemia cancer

Models	Confusion Matrix	Sensitivity	Specificity
Logit model	9020 2106 2125 5994	0.8093	0.7401
Random forest	9125 2141 2020 5959	0.8187	0.7356
Decision tree	8990 2111 2155 5989	0.8066	0.7393
SVM linear	8022 2358 3123 5742	0.7197	0.7088
SVM radial	8129 2371 3016 5729	0.7293	0.7073
SVM sigmoid	7996 2479 3149 5621	0.7174	0.6939
Neural network	9230 2128 1915 5972	0.8281	0.7373
Naïve Bayes	8945 2355 2200 5745	0.8026	0.7092
LDA	9225 2202 1920 5898	0.8277	0.7281
KNN	9120 2265 2025 5835	0.7771	0.7203

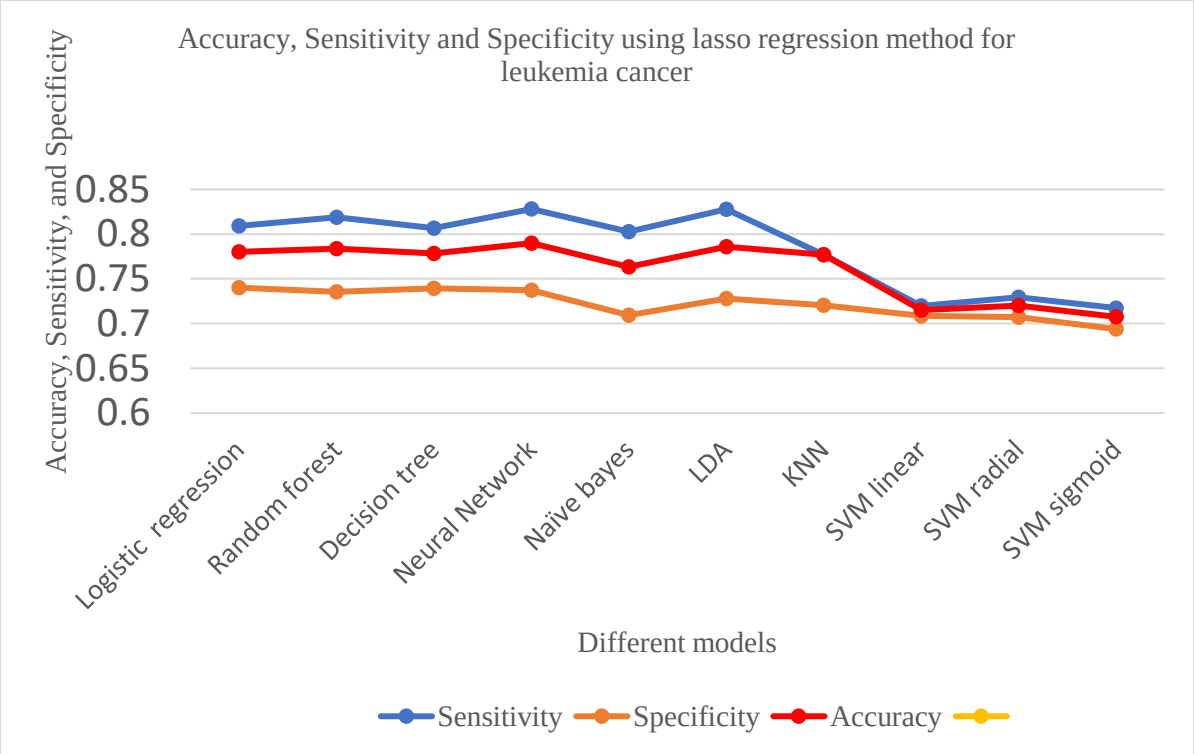

Figure 31. Comparing leukemia cancer accuracy, sensitivity, and specificity with different methods using the lasso regression method

The ANN had the highest accuracy of 78.99% followed closely by LDA (78.58%) and RF (78.38%) respectively. The SVM sigmoid model had the lowest accuracy of 70.75% (Figure 31). ANN was better than the other models with a sensitivity of 82.81% followed closely by LDA (82.77%). The sensitivity 82.81% indicates that 82.81% of correct predictions among leukemia cancer patients who survived. The SVM sigmoid model had the lowest sensitivity among all other models at almost 71.74%.

The LR had the highest specificity among all other methods at almost 74.01%. The SVM sigmoid had the lowest specificity among all other methods almost at 69.39% (Figure 31 and Table 28).

Table 29. Confusion matrix using variables selection by Boruta algorithm for leukemia cancer.

Models	Confusion Matrix		Sensitivity	Specificity
Logit model	9902	2004	0.8884	0.7526
	1243	6096		
Random forest	9965	1947	0.8941	0.7596
	1180	6153		
Decision tree	9919	2079	0.8899	0..7433
	1226	6021		
SVM linear	8976	2394	0.8054	0.7044
	2169	5706		
SVM radial	8891	2258	0.7977	0.7212
	2254	5842		
SVM sigmoid	8870	2218	0.7958	0.7262
	2275	5882		
Neural network	9932	2016	0.8911	0.7511
	1213	6084		
Naïve Bayes	9792	2128	0.8786	0.7372
	1353	5972		
LDA	9866	2111	0.8852	0.7394
	1279	5989		
KNN	9802	2139	0.8794	0.7359
	1343	5961		

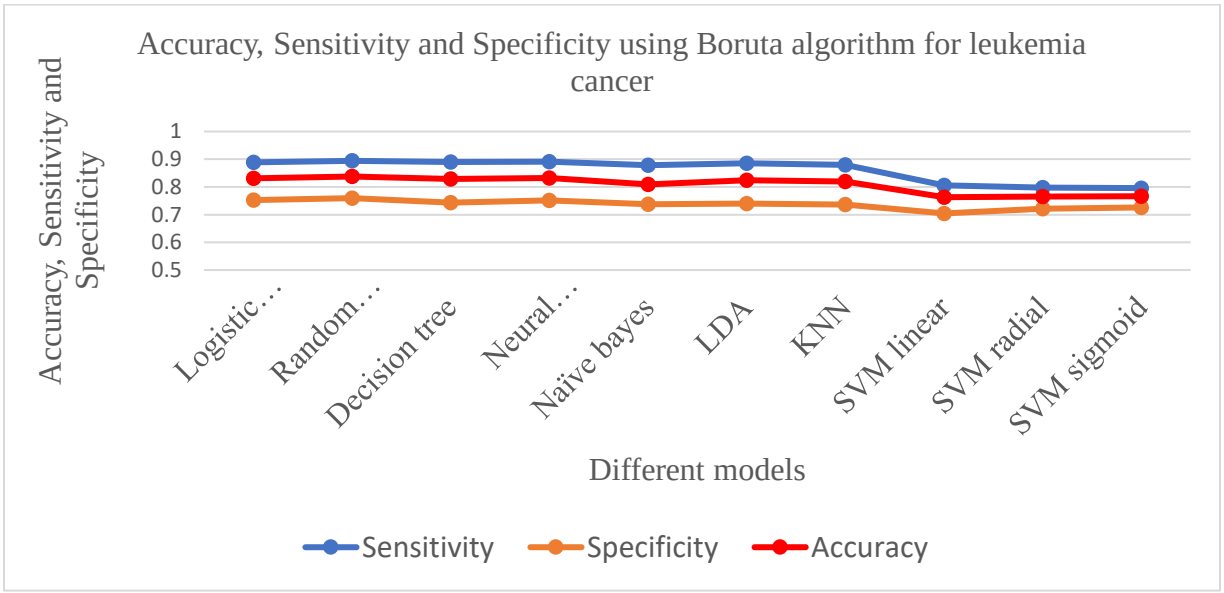

Figure 32. Comparing leukemia cancer accuracy, sensitivity, and specificity with different methods using Boruta algorithms.

The RF had the highest accuracy of 83.75% followed closely by LR (83.13%) and DT (82.83%) methods respectively. The SVM linear model had the lowest accuracy of 76.29% (Figure 32). RF was better than the other models with a sensitivity of 89.41% followed closely by ANN (89.11%), DT (88.99%), and LR (88.84%) respectively. The sensitivity of 89.41 % indicates that 89.41% of correct predictions among leukemia cancer patients who survived. The SVM radial model had the lowest sensitivity among all other models at almost 79.77%. The RF model had the highest specificity among all other methods at almost 75.96% followed closely by LR (75.26%). The SVM linear had the lowest specificity among all other methods almost at 70.44% (Figure 32 and Table 29).

7.2. Area Under the Curve for Leukemia Cancer Using ML and Data Reduction Methods

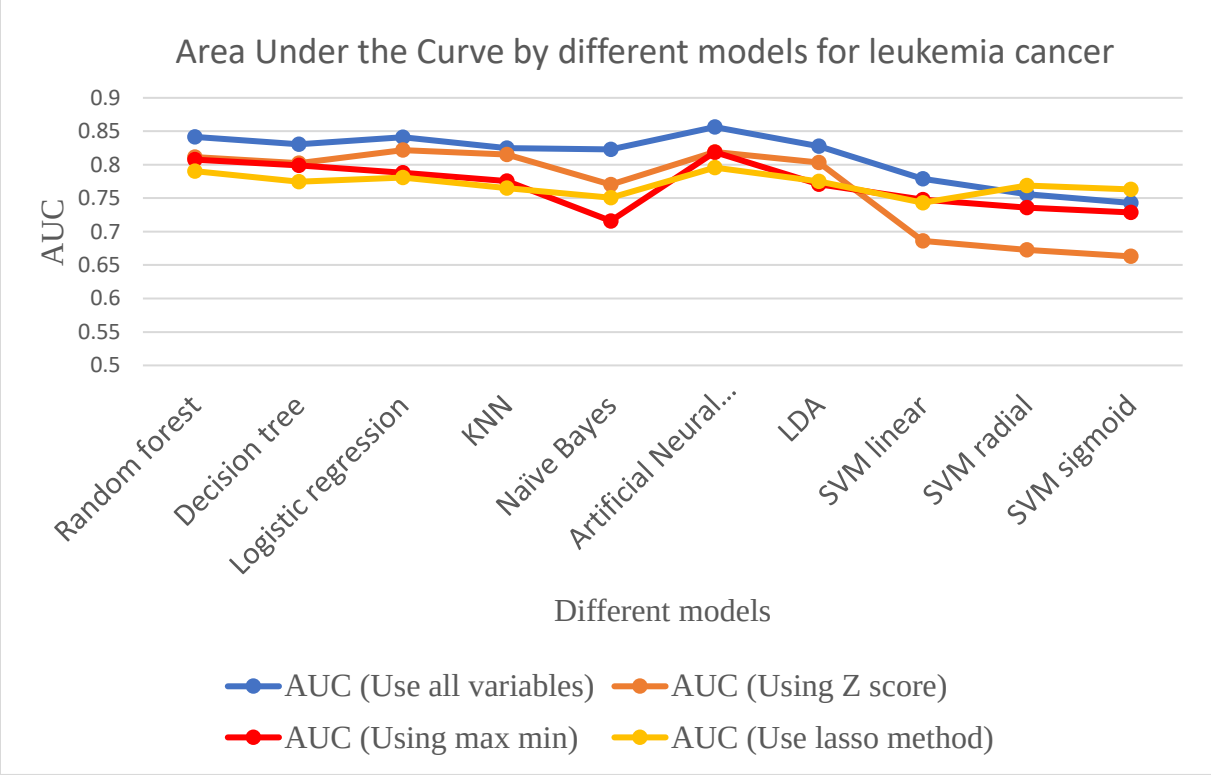

Figure 33. Compare leukemia cancer AUC with different models using different data sets

The AUC has been obtained for the various machine learning and data reduction techniques using four different datasets with different independent variables. The ANN had the highest AUC

at almost 85.63% followed closely by random forest (84.17%) using the variables selection method via the Boruta algorithm. In addition, ANN had also the highest AUC using PCs from the Z score as well as max-min normalization techniques. On the other hand, the SVM had the lowest AUC using all datasets (Figure 33).

CHAPTER VIII: CONCLUSION

In our study, we compared five-year survival prediction accuracy for lung, breast, colon, and leukemia cancers using several machine learning and data reduction techniques. We used a quite large dataset from the SEER program under the National Cancer Institute. After going through a process of data cleaning and transformation, we have developed the best survival prediction and accuracy models. Four data sets, each with different independent variables, were used to compare the model's performances of the various machine learning and data reduction techniques. Four different data sets had selected using the Boruta algorithm, Lasso regression, and PCA methods. For the PCA method, we used Z score and max-min normalization techniques to get principal components as an independent variable. Furthermore, we used a cross-validation procedure with a 10-fold to measure the unbiased different cancers' five-year survival prediction accuracy. The LDA technique had the highest breast cancer accuracy of 68.42%, followed closely by the RF (68.14%) and ANN (68.10%) methods based on the Z score.

Based on the other data reduction method max-min normalization for breast cancer, the LR technique had the highest accuracy of 68.97%, followed closely by the RF (68.48%) and ANN (68.14%) methods. In addition, the LR model also had the best model with a sensitivity of 85%. Based on the specificity, the NB had the best model with a specificity of 51.76% followed closely by the LDA (51.69%) and RF (51.65%) methods. Based on the lasso method for breast cancer, the LR, and RF both had the best survival prediction models with the highest accuracy of 70%. According to the sensitivity results, LR had the best model. On the contrary, the SVM linear (54.82%) had the best model followed closely by RF (54.27%) based on the specificity.

Based on the variable selection by the Boruta algorithm for breast cancer, the RF had the best model with an accuracy of 71.49% followed closely by the LR (71.28%) and DT (71.16%).

The sensitivity results indicated that RF also had the best model with a sensitivity of 83.87% followed closely by the DT (83.74% and LR (83.61%). Based on the specificity, the SVM radial had the best model with a specificity of 57.06 %followed closely by the SVM sigmoid (56.65%) and the SVM linear (56.39%). According to the AUC, the RF (69.16%) had the best model followed closely by the DT (68.79 %) and LR (68.94 %) using the variables selection technique via the Boruta algorithm.

As previously mentioned, Delen et.al (2005) used SEER breast cancer data for five-year survival prediction from 1973 to 2000 with three machine learning techniques including the C4.5 decision tree, ANN, and logistic regression. The authors showed that the C4.5 decision tree method was the best with an accuracy of 93.6% for breast cancer. Based on the sensitivity of 96.02%, C4.5 was the best closely followed by ANN with 94.37% sensitivity. In this study, AUC was not considered for model performance. Rajesh and Anand (2012) implemented KNN, C4.5, and NB machine learning techniques. The study showed 94% accuracy for the C4.5 model, 93% accuracy for KNN, and 92% accuracy for the NB classifier. We observed that in comparison to the results of both papers, our accuracy was much lower. There is more imbalance class in the older SEER data set (from 1973 to 2000) compared to new data set from 2004 to 2016. Delen reported that 54% did not survive but 46% survived. In that case, accuracy showed more belongs to did not survive group. It can be bias. Therefore, their accuracy showed higher. We could not find any of papers who used SEER data sets from 2004 to 2016 and their results were better than our results. We believe that in the perspective of data set from 2004 to 2016 our results were acceptable. In addition to the prediction model, we also conducted sensitivity and specificity analysis on several machine learning and data reduction models to gain the best prediction model for breast cancer five-year survivability. The sensitivity results indicated that the logistic model had the best model

with a sensitivity of 84%. On the other hand, LDA had the best model with a specificity of 53% based on the Z score data reduction method for breast cancer.

For lung cancer, the ANN model had the best model with an accuracy of 68.41% based on the Z score. It was also the best model with a sensitivity of 72.44%. The SVM linear had the best model with a specificity of 63.79%. Furthermore, the RF technique had the best model with an accuracy of 67.72 % using max-min normalization. Based on the sensitivity of 73.48%, the RF was the best model. The SVM radial had the best model with a specificity of 62.87 %. Based on the variable selection via the lasso method, the RF model was the best model with an accuracy of 68.14 %. It was also the best model based on the sensitivity of 70.46%. The LDA had the best model according to the specificity with 64.09%. The RF model for lung cancer had the best accuracy model of 71.62% using the variables selection technique from the Boruta algorithm. Based on the sensitivity and specificity, the RF method had the best model with a sensitivity of 75.75% and specificity of 66.38%. According to AUC, the RF was the best model at almost 72.58% followed closely by LR (72.08%) using the variables selection technique from the Boruta algorithm.

Agrawal et. al (2012) conducted a lung cancer survival prediction study using SEER data from 1998 to 2008 and used three machine learning techniques including decision tree, random forest, and ensemble methods. They predict 6 months, 9 months, 1 year, 2 years, and 5 years of survival prediction. Among the ML model, ensemble voting classification techniques performed the best technique with 91.35% accuracy for 5-year lung cancer survival prediction. They did not consider AUC for model performance. We observed that they used different data sets, therefore, our accuracy results were different than their accuracy results. Furthermore, higher accuracy results sometimes can be misleading due to imbalance classes. Agrawal et.al showed the imbalance

classes for five-year survival such as 83.23% not survive and 16.77% survived. In that case, accuracy showed more belongs to not survive group. It can be bias.

For colon cancer, the LR model had the best model with an accuracy of 65.72% using the Z score. The RF model had the best model with a sensitivity of 65.78% followed closely by the ANN method (64.89%). The LR had the best model with a specificity of 68.19%. On the other hand, the LR had the best model with an accuracy of 63.33% based on the max-min normalization. The RF method had the best method with a sensitivity of 63.11% followed closely by the DT method (62.59%). The LR had the best model with a specificity of 68.45%. Based on the variable selection via the lasso regression method, the ANN had the best model with an accuracy of 66.46 % followed closely by RF (66.33%), and LR (65.88%). The RF method had the best model with a sensitivity of 69.81 % followed closely by the ANN (69.51 %), and LR (69.15%). The ANN had the best model with a specificity of 63.14 % followed closely by LDA (62.60%), KNN (62.71%), and RF (62.53%). Based on all variables results, the ANN had the best model with an accuracy of 71.93% and with a sensitivity of 78.83%, The SVM linear model had the best model with a specificity of 67.32%. The ANN had the best model with an AUC of 77.15% followed closely by LR (76.11 %), and RF (76.16%) using the variables selection technique via the Boruta algorithm.

Al-Bahrani et al., (2013) reported five years of colon cancer survival prediction using SEER data from 1973 to 2009. In this study, some ML techniques were used such as RF, DT, LR, and Ensemble Voting. The highest accuracy of 90.38% for the ensemble method for five years of survival. Furthermore, the AUC for five years was 0.92 for the ensemble voting method. But they were not selected the best prediction model based on sensitivity, and specificity. Indeed, our prediction accuracy results were lower than their results because of different data sets.

For leukemia cancer, the ANN had the best model with an accuracy of 81.72 %, followed closely by the RF (81.74 %) and LR (81.13 %) using the Z score. Based on the sensitivity results, the RF was the best model with a sensitivity of 88.88 %. The ANN had the best model with a specificity of 73.14%. Based on the max-min normalization, the ANN had the best model with an accuracy of 82.42 %, a sensitivity of 88.79%, and a specificity of 73.65%. Furthermore, the ANN had the highest accuracy of 78.99 % followed closely by LDA (78.58 %) and random forest (78.38 %) based on the lasso regression method. In addition, ANN was better than the other models with a sensitivity of 82.81%. The LR had the best model based on the specificity of 74.01%. Based on the variable selection by the Boruta algorithm, the RF had the best model with an accuracy of 83.75%, a sensitivity of 89.41%, and a specificity of 75.96%. The ANN had the best model based on the AUC with 85.63%.

Hassouneh et al. (2019) showed worse accuracy compared to our results of leukemia survival. The accuracy of DNN was about 75%, DT and SVM were about 73.45%, and ANN was about 74%. Therefore, DNN was the best model rather than the three ML techniques. Finally, we can conclude that most of the previous papers used the SEER data from 1973 to 2003, therefore, their survival prediction accuracy results were different than our prediction accuracy results. In addition, most of the papers, they were not considered 10-fold cross-validation approach, sensitivity, specificity, and AUC for their results. On the other hand, we considered all machine learning techniques together as well as 10-fold cross-validation approach, sensitivity, specificity, and AUC for our results. Overall, it is apparent that ANN, RF, and LR were the best prediction models for all cancers using the Boruta algorithm.

8.1. Research Contribution

There is no study using SEER datasets from 2004 to 2016 for cancer survivability prediction using machine learning techniques. Moreover, there is no comprehensive study using statistical modeling and machine learning techniques for cancer survival prediction. A few studies used modern machine learning techniques for cancer survival prediction, but they did not consider all models together for prediction accuracy. The following research contributes to our study:

- For each cancer type, four data sets were constructed for comparison purposes, each with different predictor variables. The four data sets were: data reduction using PCs from Z score normalized data of predictor variables, data reduction using PCs from max-min normalized data of predictor variables, variables selected using lasso regression, and Boruta algorithms. The creation of these different sets of predictor variables allowed us to compare machine learning and data reduction techniques simultaneously.
- The following model combinations were used to get prediction accuracy. The combinations of models are Random Forest using PCA, Logistic using PCA, decision tree using PCA, Neural network using PCA, Naïve Bayes using PCA, LDA using PCA, and SVM using PCA. These combinations were not used for cancer survival prediction before.
- We compared AUC for all models to identify the best prediction methods.
- We used the 10-fold cross-validation method because it gives more feasible results. The dataset was randomly divided into 10 disjoint folds. Each fold contained approximately the same number of records. For each subset, a classifier is constructed using nine of the 10 folds and tested on the tenth one to obtain a cross-validation estimate of its error rate. The 10 cross-validation estimates are then averaged to provide an estimate for the classifier accuracy constructed from all the data.

- We applied all machine learning and data reduction techniques to get the best accuracy methods not only for lung cancer, but also for breast cancer, colon cancer, and leukemia as well using the SEER data.
- The accuracy of various machine learning techniques is acceptable and can help the medical professional in decision-making for early diagnosis and avoiding biopsy.

REFERENCES

ACS Journal. (n.d.). https://acsjournals.onlinelibrary.wiley.com/doi/full/10.3322/caac.21590

American Cancer Society. (n.d.). https://www.cancer.org

American Cancer Society (n.d) https://www.cancer.gov/about-cancer/causes-prevention/risk/tobacco

Abdelaal, M. M. A., Sena, H. A., Farouq, M. W., & Salem, A. M. (2010). Using Data Mining for Assessing Diagnosis of Breast Cancer. *Proceedings of the International Multiconference on Computer Science and Information Technology*, 11–17. https://doi.org/10.1109/IMCSIT.2010.5679647

Agrawal, A., Misra, S., Narayanan, R., Polepeddi, L., & Choudhary, A. (2012). Lung Cancer Survival Prediction using Ensemble Data Mining on SEER Data. *Scientific Programming*, *20*, 29–42. https://doi.org/10.3233/SPR-2012-0335

Al-Bahrani, R., Agrawal, A., & Choudhary, A. (2013). Colon cancer survival prediction using ensemble data mining on SEER data. *2013 IEEE International Conference on Big Data*, 9–16.

Anand, P., Kunnumakkara, A. B., Sundaram, C., Harikumar, K. B., Tharakan, S. T., Lai, O. S., Sung, B., & Aggarwal, B. B. (2008). Cancer is a Preventable Disease that Requires Major Lifestyle Changes. *Pharmaceutical Research*, *25*(9), 2097–2116. https://doi.org/10.1007/s11095-008-9661-9

Bartholomai, J. A., & Frieboes, H. B. (2018). Lung Cancer Survival Prediction via Machine Learning Regression, Classification, and Statistical Techniques. *2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, 632–637. https://doi.org/10.1109/ISSPIT.2018.8642753

Bellaachia, A., & Guven, E. (2006). Predicting breast cancer survivability using data mining techniques. *Age*, *58*(13), 10–110.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, 144–152.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Burke, H. B., Goodman, P. H., Rosen, D. B., Henson, D. E., Weinstein, J. N., Harrell Jr., F. E., Marks, J. R., Winchester, D. P., & Bostwick, D. G. (1997). Artificial Neural Networks Improve the Accuracy of Cancer Survival Prediction. *Cancer*, *79*(4), 857–862. https://doi.org/https://doi.org/10.1002/(SICI)1097-0142(19970215)79:4<857::AID-CNCR24>3.0.CO;2-Y

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, *20*(3), 273–297.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*(1), 21–27.

Delen, D., Walker, G., & Kadam, A. (2005). Predicting Breast Cancer Survivability: a Comparison of Three Data Mining Methods. *Artificial Intelligence in Medicine*, *34*(2), 113–127. https://doi.org/https://doi.org/10.1016/j.artmed.2004.07.002

Doll, R., Peto, R., Boreham, J., & Sutherland, I. (2004). Mortality in relation to smoking: 50 years' observations on male British doctors. *Bmj*, *328*(7455), 1519.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, *7*(2), 179–188.

Fradkin, D., Schneider, D., & Muchnik, I. (2006). Machine Learning Methods in the Analysis of Lung Cancer Survival Data. *DIMACS Technical Report 2005–35*.

Gadgeel, S. M., & Kalemkerian, G. P. (2003). Racial Differences in Lung Cancer. *Cancer and Metastasis Reviews*, *22*(1), 39–46.

Gareth, J., Daniela, W., Trevor, H., & Robert, T. (2013). *An introduction to statistical Learning: with Applications in R*. Spinger.

Ganggayah, M. D., Taib, N. A., Har, Y. C., Lio, P., & Dhillon, S. K. (2019). Predicting Factors for Survival of Breast Cancer Patients using Machine Learning Techniques. *BMC Medical Informatics and Decision Making*, *19*(1), 48. https://doi.org/10.1186/s12911-019-0801-4

Genuer, R., Poggi, J.-M., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, *31*(14), 2225–2236.

Gultepe, Y. (2021). Performance of Lung Cancer Prediction Methods Using Different Classification Algorithms

Han, H., Guo, X., & Yu, H. (2016). Variable Selection Using Mean Decrease Accuracy and Mean Decrease Gini Based on Random Forest. *2016 7th Ieee International Conference on Software Engineering and Service Science (Icsess)*, 219–224.

Hankey, B. F., Ries, L. A., & Edwards, B. K. (1999). The surveillance, Epidemiology, and End Results Program: a National Resource. *Cancer Epidemiology and Prevention Biomarkers*, *8*(12), 1117–1121.

Hassouneh, N., Alnemer, L., Alsakran, J., & Rodan, A. (2019). Predicting Survivability in Leukemia Patients using Deep Learning. *2019 Sixth HCT Information Technology Trends (ITT)*, 191–196.

Hastie, T., & Qian, J. (2014). Glmnet vignette. *Retrieved June*, *9*(2016), 1–30.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2). Springer.

Heath, M., Bowyer, K., Kopans, D., Kegelmeyer, P., Moore, R., Chang, K., & Munishkumaran, S. (1998). Current status of the digital database for screening mammography. In *Digital mammography* (pp. 457–460). Springer.

Ho, T. K. (1995). Random decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition*, *1*, 278–282.

Hong, Z.-Q., & Yang, J.-Y. (1991). Optimal discriminant plane for a small number of samples and design method of classifier on the plane. *Pattern Recognition*, *24*(4), 317–324.

Hosmer, D. W., Lemeshow, S., & Lemeshow, S. (2000). *Applied Logistic Regression: Wiley Series in Probability and Statistics: Texts and References Section*. Wiley Hoboken, NJ, USA.

Jajroudi, M., Baniasadi, T., Kamkar, L., Arbabi, F., Sanei, M., & Ahmadzade, M. (2014). Prediction of Survival in Thyroid Cancer Using Data Mining Technique. *Technology in Cancer Research & Treatment*, *13*(4), 353–359. https://doi.org/10.7785/tcrt.2012.500384

Kursa, M. B., & Rudnicki, W. R. (2010). Feature selection with the Boruta package. *J Stat Softw*, *36*(11), 1–13.

Liou, D. M., & Chang, W. P. (2015). Applying data mining for the analysis of breast cancer data. *Methods in Molecular Biology*, *1246*(7455), 175–189. https://doi.org/10.1007/978-1-4939-1985-7_12

Lundin, M., Lundin, J., Burke, H. B., Toikkanen, S., Pylkkänen, L., & Joensuu, H. (1999). Artificial Neural Networks Applied to Survival Prediction in Breast Cancer. *Oncology*, *57*(4), 281–286. https://doi.org/10.1159/000012061

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5*(4), 115–133.

Ming, C., Viassolo, V., Probst-Hensch, N., Chappuis, P. O., Dinov, I. D., & Katapodi, M. C. (2019). Machine Learning Techniques for Personalized Breast Cancer Risk Prediction: Comparison with the BCRAT and BOADICEA Models. *Breast Cancer Research*, *21*(1), 75. https://doi.org/10.1186/s13058-019-1158-4

Minsky, M., & Papert, S. (1969). *Perceptron: an introduction to computational geometry*. Cambridge, MA: MIT Press.

Mourad, M., Moubayed, S., Dezube, A., Mourad, Y., Park, K., Torreblanca-Zanca, A., Torrecilla, J. S., Cancilla, J. C., & Wang, J. (2020). Machine Learning and Feature Selection Applied to SEER Data to Reliably Assess Thyroid Cancer Prognosis. *Scientific Reports*, *10*(1), 5176. https://doi.org/10.1038/s41598-020-62023-w

Rajesh, K., & Anand, S. (2012). Analysis of SEER Dataset for Breast Cancer Diagnosis using C4. 5 Classification Algorithm. *International Journal of Advanced Research in Computer and Communication Engineering*, *1*(2), 1021–2278.

Salod, Z., & Singh, Y. (2019). Comparison of the Performance of Machine Learning Algorithms in Breast Cancer Screening and Detection: A protocol. *Journal of Public Health Research*, *8*(3), 1677. https://doi.org/10.4081/jphr.2019.1677

Santosa, F., & Symes, W. W. (1986). Linear inversion of band-limited reflection seismograms. *SIAM Journal on Scientific and Statistical Computing*, *7*(4), 1307–1330.

Sharma, R., Pannikottu, J., Xu, Y., Tung, M., Nothelle, S., Oakes, A. H., & Segal, J. B. (2018). Factors Influencing Overuse of Breast Cancer Screening: A Systematic Review. *Journal of Women's Health (2002)*, *27*(9), 1142–1151. https://doi.org/10.1089/jwh.2017.6689

Shelly Gupta, Dharminder Kumar, A. S. (2011). Data Mining Classification Techniques Applied for Breast Cancer Diagnosis and Prognosis. *Indian Journal of Computer Science and Engineering*, *2*(2).

Siegel, R. L., Miller, K. D., Goding Sauer, A., Fedewa, S. A., Butterly, L. F., Anderson, J. C., Cercek, A., Smith, R. A., & Jemal, A. (2020). Colorectal cancer statistics, 2020. *CA: A Cancer Journal for Clinicians*, *70*(3), 145–164.

Street, W. N., Wolberg, W. H., & Mangasarian, O. L. (1993). Nuclear feature extraction for breast tumor diagnosis. *Biomedical Image Processing and Biomedical Visualization*, *1905*, 861–870.

Tan, P.-N., Steinbach, M., & Kumar, V. (2016). *Introduction to Data Mining*. Pearson Education India.

Tibshirani, R. J. (2011). *Regression shrinkage and selection via the lasso*.

Yeh, W.-C. (2019). A novel generalized artificial neural network for mining two-class datasets. *ArXiv Preprint ArXiv:1910.10461*.

APPENDIX. LIST OF VARIABLES AND VARIABLES DESCRIPTIONS

Marital Status	
Category	Description
M1	Marital Status single
M2	Marital Status Married
M3	Marital Status Separated
M4	Divorced
M5	Widowed
M6	Unmarried or Domestic Partner
Race	
R1	Race White
R2	Race Black
R3	American Indian
R4	Chinese
R5	Japanese
R6	Filipino
R7	Hawaiian
R8	Korean
R9	Vietnamese
R10	Laotian
R11	Hmong
R12	Kampuchean
R13	Asian Indian
R14	Thai
R15	Pakistani
R16	Micronesian
R17	Chamorran
R18	Gumanian
R19	New Guinean
R20	Polynesian
R21	Tahitian
R22	Samoan
R23	Tongan
R24	Melanesian
R25	Fiji Islander
R26	Other Asian
R27	Pacific Islander
R28	Other
Stage	
S1	In situ
S2	Localized
S3	Regional, regional lymph nodes only
S4	Regional, extension and nodes
S5	Localized lymph nodes

Histoloy	
H1	Epithelial neoplasms
H2	Squamous cell neoplasm
H3	Basal cell neoplams
H4	Transitional cell papillomas and carcinomas
H5	Adenomas and adenocarcinomas
H6	Adnexal and appendage neoplasm
H7	Mucoepidermoid neoplasms
H8	Cystic, mucinous and serous neoplasm
H9	Ductal and lobular neoplasm
H10	Acinar cell neoplasms
H11	Adenocarcinomas
H12	Thymic epithelial neoplasm
H13	Specialized gonadal neoplasms
H14	Paragangliomas and glumus tumors
H15	Carcinomas
H16	Soft tissue tumors and sarcomas
H17	Fibromatous neoplasms
H18	Myxomatous neoplasms
H19	Lipomatous neoplasms
H20	Adnexal
H21	Complex mixed and stromal neoplasms
H22	Fibroepithelial neoplasms
H23	Synovial like neoplasms
H24	Lobular
H25	Neoplasm
H26	Mucinous
H27	Gonadal
H28	Adenomas
H29	Stromal
H30	Sarcomas
H31	Glioblastoma
H32	Bone tumors
H33	Ependymoma
H34	Mixed glioma
H35	Gliomas
H36	Pilocytic
H37	Meningiomas
H38	Nerve sheath tumors
H39	Oligodendroglima
H40	Malignant lymphomas
H41	Neuroepithelial
H42	Embryonal
H43	Medulloblastoma
H44	Ampulla vater

H45	Plasma cell tumors
H46	Germ cell tumors
H47	Chondrosarcoma
H48	chordoma
H49	Craniopharyngioma
H50	Lymphoid leukemias
H51	Myeloid leukemias
H52	Myelodysplastic syndrome
H53	Chronic myeloproliferative disorders
H54	Lymphona
H55	Hemagioblastoma
H56	Heterotopias
H57	Astrocytoma
H58	Appendage
H59	Anaplastic
H60	Choroid Plexus
H61	Glial
H62	Carcinoid
H63	Retroperitoneum
H64	Peritoneum
H65	Retinoblastoma
H66	Ductal
H67	Papillomas
H68	Mucinous
H69	serous
H70	Peritoneum
H71	Benign and malignant neuronal
H72	Analplastic astrocytoma
H73	Pilocytic astrocytoma
H74	Neuroepithelial
H75	Glioma, Nos
H76	Primitive plexus
H77	Medulloblastoma
H78	Pineal parenchymal
H79	BilleDusctsIntraHepat
H80	Plasma cell tumor
H81	Lymphomas
H82	Hematologic disorders
H83	Precursor lymphoblastic
H84	Mast cell tumors
H85	Unique astrocytoma
H86	Biliary other
H87	Corpus Sarcoma
H88	Epiglotttis Anterior
H89	Esophagus

H90	Squamous neoplasms
H91	Transitional carcinomas
Radiation	
Radi1	External beam
Radi2	External beam photons
Radi3	External beam electrons
Radi4	Brachytherphy, Intracavitary, LDR
Radi5	Radioisotopes, Radium
Radi6	Radioisotopes, Strontium
Radi7	Radiation therapy before surgery
Radi8	Radiation therapy after surgery
Radi9	Radiation therapy both before and after surgery
Radi10	Intraoperative radiation therapy
Grade	
g1	Grade I , well differentiated cell
g2	Grade II, moderately differentiated cell
g3	Grade III, poorly differentiated
g4	Grade IV, anaplastic
g5	Grade V, T-cell, T precursor
Behavior	
B1	Carcinoma in situ
B2	Malignant
Primary site	
Prim1	C445-Anal margin
Prim2	C221- Anal Verge
Prim3	C162
Prim4	C163-Angular incisura
Prim5	C44-Cutaneous leiomyosarcoma
Prim6	C720-Distal Conus
Prim7	C109- Glossotonsillar sulcus
Prim8	C349-Infrahilar area of lung
Prim9	C269-Pancreatobiliary
Lymnode Involvement	
Linv1	Regional lymph node involves
Linv2	Aspiration of regional lymph node
Linv3	More regional lymph nodes involves
Linv4	Sentinel node biopsy involves at same time
Linv5	Regional lymph node documented as sampling
Linv6	Regional lymph node documented as dissection
Linv7	Sentinel node biopsy involves at different time
Linv8	Biopsy of regional lymph node

Linv9	Biopsy as lymph node sampling
Linv10	No regional lymph nodes removed
Surgery	
Surg1	Surgery performed
Surg2	Surgery not recommended
Surg3	Autopsy only case
Surg4	Patient died before surgery
Surg5	Patient died after surgery
Surg6	Patient's or Patient's guardian refuse surgery
Extent of Diseases (EOD)	
E1	Autopsy only
E2	No pathologic specimen
E3	Hematopoietic
E4	Reticuloendothelial
E5	Immunoproliferative
E6	Myeloproliferative
E7	Resection
E8	EOD regional nodes
E9	EOD primary tumor
E10	Kaposi sarcoma
E11	Mycosis Fungoides
E12	Adenoma
E13	Polyp
E14	Epithelium
E15	Paraaortic lymph nodes
E16	Intraepithelial tubal mucosa
E17	Larynx Glottic
E18	Maxillary Sinus
E19	Ethmoid Sinus
E20	Merkel cell
E21	Oropharynx
E22	Pleura Mesothelioma
E23	Cervix Sarcoma
E24	Lymph node excision
E25	Buccal Mucosa
E26	Cervical lymph nodes
E27	Regional nodes positive
E28	Esophagus excluding squamous
E29	Lymphoma
Continuous variables	
AG	AGE
nplymnode	Number of lymnode positive
nlymonode	Number of regional lymnode
Tumor	Tumor size