EVALUATION OF MACHINE LEARNING ALGORITHMS FOR PREDICTION OF 5-YEAR SURVIVABILITY FOR SEVEN TYPES OF CANCERS

A Thesis presented to
the Faculty of the Graduate School
at the University of Missouri

In Partial Fulfillment
Of the Requirements for the Degree
Master of Science

By
KOTAPATI TEJA VENKAT PAVAN
Dr. ABU SALEH MOHAMMAD MOSA, Thesis Supervisor
JULY 2020

The undersigned appointed by the Dean of the Graduate School, have examined the dissertation entitled:

**EVALUATION OF MACHINE LEARNING ALGORITHMS
FOR PREDICTION OF 5-YEAR SURVIVABILITY FOR
SEVEN TYPES OF CANCERS**

presented by Kotapati Teja Venkat Pavan,
a candidate for the degree of Master of Science and hereby certify that, in their
opinion, it is worthy of acceptance.

Dr. Abu Mosa

Dr. Sue Boren

Dr. Christos Papageorgiou

DEDICATION

I dedicate this thesis to my family and my professor Dr. Abu Mosa for helping me to give the life I love today.

ACKNOWLEDGEMENTS

First, I would like to express gratitude to my thesis advisor Dr. Abu Mosa for his guidance and support throughout this research project.

I would like to acknowledge Dr. Abu Mosa, Dr. Sue Boren and Dr. Christos Papageorgiou for serving on my thesis committee and spending their valuable time.

I would also like to thank Vasanthi Mandhadi (School of Medicine's Sr. Research Design Engineer) and all of my colleagues in the Research Informatics Lab for their encouragement which inspired me in completing the project.

Finally, I would like to thank my family, friends and teachers, whose support had inspired me to make this project possible.

TABLE OF CONTENTS

Organization of the Dissertation	ix
Abstract	x
Chapter 1 : Introduction.....	1
1.1 Cancer Introduction	1
1.2 Significance of Cancer.....	2
1.3 Cancer Survivability	3
1.4 Knowledge Discovery in Database	4
1.5 Problem Statement	5
1.6 Aim	6
Chapter 2 : Background.....	7
2.1 Cancer Research	7
2.2 Previous Research	7
Chapter 3 : Methods	14
3.1 Data source.....	14
3.2 Data Dictionary	15
3.3 Data Preprocessing	16
3.3.1 Data Extraction.....	17
3.3.2 Variable Selection	17
3.3.3 Merging of Identical Variables	24
3.3.4 Data Mapping.....	25
3.3.5 Dependent Variable Creation.....	28
3.3.6 Record Elimination with Clinically Irrelevant Values	31
3.3.7 Record Elimination with Missing Values	32
3.3.8 Creation of Sparse Format:	33
3.4 Experimental Approach	34
3.5 Cross-Validation.....	36
3.6 Metrics	37
3.6.1 Sensitivity	37
3.6.2 Accuracy	37
3.6.3 Specificity	37
3.6.4 Precision	37
3.6.5 F1 Score	38

3.6.6	AUC Score.....	38
Chapter 4	: Class Balancing Techniques	39
4.1	Over Sampling	40
4.2	Under Sampling.....	40
Chapter 5	: Results	42
5.1	Breast Cancer	42
5.2	Colon & Rectal Cancer	44
5.3	Digestive & Thoracic Cancer.....	46
5.4	Male Genital Cancer	48
5.5	Female Genital Cancer	49
5.6	Respiratory Cancer	51
5.7	Urinary Cancer	53
5.8	Discussion	58
5.9	Limitations	59
5.10	Future Work	60
Bibliography.....		61

LIST OF TABLES

Table 1: Eliminated Deprecated Variables	18
Table 2: Eliminated Duplicate Variables	21
Table 3: Eliminated Irrelevant Variables	22
Table 4: Variables Chosen for Prediction of Cancer Survivability	23
Table 5: Coding Schema for Cancer Staging Variable.....	28
Table 6: Cohort Population	33
Table 7: Imbalanced Cohorts	39
Table 8: Random Over Sampling of Training Data.	40
Table 9: Random Under Sampling of Training Data.	41
Table 10: Breast Cancer No Sampling Results	43
Table 11: Breast Cancer Over-Sampling Results.....	44
Table 12: Breast Cancer Under Sampling Results	44
Table 13: Colon & Rectal Cancer No Sampling Results.....	45
Table 14: Colon & Rectal Cancer Over Sampling Results.....	45
Table 15: Colon & Rectal Cancer Under Sampling Results.....	46
Table 16: Digestive & Thoracic Cancer No Sampling Results	47
Table 17: Digestive & Thoracic Cancer Over Sampling Results.....	47
Table 18: Digestive & Thoracic Cancer Under Sampling Results.....	47
Table 19: Male Genital Cancer No Sampling Results.....	48
Table 20: Male Genital Cancer Over Sampling Results.....	49
Table 21: Male Genital Cancer Under Sampling Results.....	49
Table 22: Female Genital Cancer No Sampling Results	50

Table 23: Female Genital Cancer Over Sampling Results	50
Table 24: Female Genital Cancer Under Sampling Results	51
Table 25: Respiratory Cancer No Sampling Results	52
Table 26: Respiratory Cancer Over Sampling Results	52
Table 27: Respiratory Cancer Under Sampling Results	52
Table 28: Urinary Cancer No Sampling Results	53
Table 29: Urinary Cancer Over Sampling Results	54
Table 30: Urinary Cancer Under Sampling Results	54

LIST OF FIGURES

Figure 1: Data Preprocessing Steps	17
Figure 2: Variable Selection Process	24
Figure 3: Merging of Tumor Size Variable	25
Figure 4: Flow Diagram of Creation of Dependent Variable	30
Figure 5: Creation of Sparse Format	34
Figure 6: Machine Learning Experimental Approach	35
Figure 7: Cross Validation (Green Color: Training Data, Blue Color: Testing Data)	36
Figure 8: F1 Score Heatmap for No Sampling Technique	55
Figure 9: F1 Score Heatmap for Over Sampling Technique	55
Figure 10: F1 Score Heatmap for Under Sampling Technique.....	56
Figure 11: AUC Heatmap for No Sampling Technique	56
Figure 12: AUC Heatmap for Over Sampling Technique	57
Figure 13: AUC Heatmap for Under Sampling Technique	57

LIST OF EQUATIONS

Equation 1: Cross Validation Accuracy (CVA)	36
---	----

ORGANIZATION OF THE DISSERTATION

Chapter 1 presents introduction, background on cancer, focus of the project and the aims of the project. **Chapter 2** presents the previous research done for predicting five-year cancer survivability using machine learning models. **Chapter 3** presents the data pre-processing steps. **Chapter 4** presents class balancing techniques. **Chapter 5** presents results on the best predictor of the machine learning models, discussion and future work.

ABSTRACT

A lot of research on prediction of cancer survivability has been done by implementing various machine learning models and it has always been a challenging task. In this project, the main focus is to perform a comprehensive evaluation of machine learning models across multiple cancer cohorts and find the models with better prediction capability. Class balancing techniques like oversampling and undersampling were implemented into the models to improve the performance of cancer survival prediction. SEER cancer dataset (1973-2015) was used for this project. After preprocessing, we included a total of 21 independent variables and a dependent variable. Multiple machine learning models like Decision Trees, Logistic Regression, Naive Bayes, Support Vector Machine, Random Forests and Multi-Layer Perceptron were implemented. Bias between training and testing data was eliminated by implementing stratified 10-fold cross-validation. The experimental design was in such a way that all the machine learning models were implemented across seven cancer cohorts using all eligible records each cohort as well as using two sampling techniques for class balancing. Performance of the machine learning models were compared based on the metrics like Sensitivity, Accuracy, Specificity, Precision, F1 score and AUC scores. A total of 168 experimental models were designed and implemented. Comparison between the predictive models showed that Random Forests have best predicted for cancer survivability, Support Vector Machine came as second-best predictors, Logistic Regression as third, then Decision Trees, Multi-Layer Perceptron and lastly Naive Bayes with least performance. The results clearly indicated that implementing class balancing techniques also improved the performance of the models significantly.

CHAPTER 1 : INTRODUCTION

1.1 CANCER INTRODUCTION

Cancer can be defined as heterogenous disease with uncontrolled cell division and capability to spread to the surrounding tissue or body, there are more than 200 different types of cancers currently detected in humans. It has the characteristics like continuous cell division, no programmed cell death, invasion of tissues and promotion of blood vessel growth at the site of occurrence. It can develop at any part of the body. Normally human cells grow, divide as a requirement to the body and as they grow old, these cells might be damaged or die naturally as programmed, this is called as apoptosis. In case of cancer cells, apoptosis does not occur, and new cells form even when they are not needed. These cells increase in number and finally form a mass of cells called tumor. Generally, cancer tumors can be categorized as two types. One of them is benign, which is not harmful because it can be removed surgically, and they cannot regrow. But once the tumor gets malignant or metastasized, they grow and spread to other parts of the body.

Cancer is a genetic disease which is caused by changes in genes (mutation), they can arise due to damaged DNA caused by certain environmental exposures or factors. They

include chemicals like tobacco and benzene compounds, radiation like UV rays and X-rays, diet, hormones etc. Cancers can be treated by procedures like chemotherapy, radiation and surgery [1]. All these treatments depend on the type of cancer, stage of cancer and the location where the cancer is present. Currently, oncologists prefer combination of these treatments to effectively treat cancer.

1.2 SIGNIFICANCE OF CANCER

Cancer has a major impact on the society, In the year 2020, National Cancer Institute estimated that 1.8 million people in the United States will be diagnosed with cancer and more than 600,000 people would die. Breast cancer is the most diagnosed cancer with more than 275,000 cases making it most common diagnosis, Lung and Bronchus cancer is the second most commonly diagnosed cancer with an estimated 230,000 cases. Prostate cancer currently is the leading cause of cancer in men and third most commonly diagnosed with overall 190,000 cases expected in 2020. Among the diagnosed cancers, Breast cancer has the highest survival rate of more than 90%, Lung and Bronchus has most deaths with 135,000 within the span of 60 months/5 years and having less than 20% of survival rate, and second most deaths occur in Colorectal cancer. Currently, the rate of new cancers of all sites is 442.4 per 100,000 in both men and women per year based on years 2013-2017 and death rate was approximately 158 per 100,000 persons per year. Over the course of time, the survival rate has increased. Currently the five-year survival rate is 67.4% on all cancer sites [2]. With increased prominence towards cancer research, new and innovative developments towards early diagnosis and treatment have been developed. It has been shown that overall cancer incidence rates are higher

among men than women. Cancer is even related to race of the person, statistics have proven that there are more cases among African American men and white women, fewer cases among Asians in both sexes. Cancer related deaths have been higher among middle aged and elderly people, median age of death is found to be 72 years. Over the last 10 years, by using the statistical analysis, age-adjusted rates for new cancer of any site, cases have been reducing by 1.1% each year.

1.3 CANCER SURVIVABILITY

Cancer survivability can be defined as the time period a patient remaining alive after being diagnosed. Cancer survival rates can be helpful for oncologists and other medical personnel to understand the prognosis and develop a treatment plan for any particular patient. Survival of a patient can be determined by the conditions like age, race, site of cancer, stage of cancer, grade of cancer, number of tumors, type of cancer and other conditions [3–5]. So, based on these conditions which are documented and along with those many number of historical records provide evidence strongly to predict the survival outcome of a particular patient.

In United States CDC's Division of cancer led a special issue on cancer survival. The populations for the study were grouped by race, stage in which the patient lived when diagnosed with cancer, how far the cancer has spread and the type of cancer. There is also a cancer survival tracking program around the world called CONCORD [6]. It provided evidence to scientists on how patients with certain cancer live longer in some countries than in other. Cancer survival studies in US have provided a lot of key findings, like 80% of women with breast cancer, men with prostate cancer and children with acute

lymphoblastic leukemia lived at least 5 years after the diagnosis. They also found a large difference in survival time between African American and white people.

In our study we are using SEER data to predict cancer survivability. In general cancer five-year survivability can be explained with an example. Let's consider a cancer with survival rate more than 90% i.e. with a population of 100 that had been diagnosed with cancer, 90 of them might survive more than 5-years and the remaining 10 do not.

1.4 KNOWLEDGE DISCOVERY IN DATABASE

Cancer research is mostly clinical or biological in nature, it mostly includes drug testing clinical trials on cancer patients and deducing the results from them. In recent years, data driven research has become more interesting in the medical domain where a lot of data is untouched. Basic statistical models are used to gain a few insights and identify further research directions for clinical and biological research. The results of data driven research have provided a strong statistical evidence that data obtained can be tied to genetic research i.e. connecting it to genes. In the present-day situation, the amount of data collected is too vast and stored in databases in medical fields and other areas. Earlier days, many statistical tools were implemented to gain new information from the data. But as the data has increased both quantitatively and qualitatively, the traditional statistical techniques have become complicated to perform the necessary data analysis. Due to this, new advanced techniques have come into light, a major area called Knowledge Discovery in Database (KDD) which is described as the process of identifying valid, novel, potentially useful and ultimately hidden understandable patterns in data and implementing machine learning techniques [7]. KDD is a process of steps in understanding the domain, understanding the

data related to domain, data preparation which includes variable merging, mapping the variables, clearing out the noise from the raw data, dealing with the missing values by various methods, creation of target variable etc.

The steps of gathering the data, gaining knowledge from it and using the pattern recognition techniques for specific purposes also comes under the area of data mining. It can be considered as a central step in the overall of KDD and due to the centrality of data mining, both the terms can be regarded as synonymous [8]. The term data mining can be described as the process of sifting through large datasets in search of interesting hidden patterns. It provides the tools to analyze the large quantity of data; due to this some prefer the term “Knowledge Mining”.

1.5 PROBLEM STATEMENT

Cancer has a major impact on society, more than 600,000 people die due to cancer every year across the USA. Due to various advancements and development in the area of cancer, early detection and effective treatments have improved the survival outcome over the years [9]. Cancer survivability helps physicians to understand the prognosis and develop treatment plans to the patients. Over the years, physicians have used historical data to estimate the survival time period, later on statistical techniques have been developed to calculate the survival time period based on the patient conditions. Recent advancements in the areas of Knowledge Discovery in Database (KDD), machine learning techniques have proven to be more accurate to predict the survivability of cancer patients based on various factors. In the current study, based on the historical data, processing the data and choosing

the right variables, we chose to predict the Five-year survival outcome on multiple cancer cohorts more accurately.

1.6 AIM

The main aim of the project is to use data mining techniques and build a prediction model which best predicts the five-year cancer survivability across multiple cancer types and at the same time implement techniques to improve the performance of these models.

CHAPTER 2 : BACKGROUND

2.1 CANCER RESEARCH

Most of the cancer research is based on pre-clinical & clinical trials, observational studies. In case of Preclinical and clinical trials, they include testing of drugs on animals and humans. These clinical trials show the cause and effect between variables and the outcome. These trials mostly have been used in developing the drugs and procedures to treat cancers. In the current study, we are going by data driven approach to implement the advanced techniques to handle large amounts of data for the prediction of survivability of cancer.

2.2 PREVIOUS RESEARCH

In this section, we will be talking about the previous research conducted in prediction of cancer survivability on various cancer types.

1. Prediction of Breast Cancer Survivability: A Comparison of Three Data Mining Techniques [10]

This article presented the severity of breast cancer and the traditional techniques that were used to predict the survivability, it also presented the advanced techniques in Knowledge Discovery and Database (KDD) which were used to create machine learning models to predict the five-year survivability of breast cancer. In this research article, they

have implemented two data mining techniques like decision tree and multi-layer perceptron, and a statistical technique called logistic regression. The machine learning models were implemented along with 10-fold cross validation, to remove the bias between the testing and training datasets. All the three models were compared to find out the best predictor. The results stated that decision tree is the best predictor for cancer survivability with an accuracy of 93.6%, multi-layer perceptron came out to be second with the accuracy of 91.2% and logistic regression is the least predictor with the accuracy of 89.2%.

2. Missing Data Imputation on the 5-Year Survival Prediction of Breast Cancer Patients with Unknown Discrete Values [11]

In this research article, prediction of five-year survivability of breast cancer by using the machine learning models like K-Nearest Neighbors, Support Vector Machines, Classification Trees and Logistic Regression. They also implemented 10-fold cross validation in the experimental setup. The main focus of this research article is to test out the better imputation technique for the missing data. So they have performed it in four scenarios, (a) Five-year survival prediction without using any data imputation technique and five year survival prediction with the clean data, (b) Mode imputation technique, where the missing data was imputed with the most repeated value, (c) Expectation-Maximization (EM) data imputation technique, and (d) KNN imputation technique. All these techniques were run on the machine learning models to predict the five-year breast cancer survivability and to find out the best imputation technique that can be used. The results concluded that, KNN classification technique using KNN imputation technique has given out the best prediction values with an accuracy of 81.73% and AUC of 0.7845. KNN imputation

technique turned out to give better results than other imputation techniques in all other classification models.

3. Breast Cancer Data Analysis for Survivability Prediction Studies and Prediction [12]

This research article has taken a different approach to predict breast cancer survivability, the main focus is to develop an analytical technique that helps in understanding the survivability in case of missing values and also providing information on the factors that are affecting the survivability of the patient. The paper talks about unsupervised learning technique to categorize the data with similar properties into cohorts, Techniques like Self-Organizing Maps (SOM) and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) were used to form the patient cohort clusters. The clustering technique was performed after data preprocessing. These clusters were then used to train the supervised MLP. A total of nine clusters were identified and they found that these have different survivability times. The results indicated that this approach i.e. using SOM and DBSCAN through attributes can explain the different survival times of the clusters.

4. Colon Cancer Survival Prediction Using Ensemble Data Mining on SEER Data [13]

The main objective of this article is to develop a survival prediction model using SEER data. In this paper, the authors used WEKA tool kit for building the data mining models, a total of 20 classification schemes were used, out of which 5 of them are basic classifiers and the remaining are a combination of three meta classifiers along with five basic classifiers. Each of these classification schemes were used to predict the survivability

for a period of one year, two years and for five years. Careful data preprocessing lead to elimination of multiple variables and finally left with a total of 65 variables. Implementing the feature selection techniques gave 13 attributes. Synthetic Minority Over-Sampling Technique (SMOTE) was used for balancing both survival and non-survival cases and the prediction models were run using the 13 attributes to check if the performance of the model is improved. The results indicated that using the class balancing technique has improved the performance of the model and ensemble voting technique has outperformed all the other classification techniques and has an overall accuracy of 90.38% for one year, 88.01% for 2 years and 85.13% for 5 years.

5. Lung Cancer Survival Prediction Using Ensemble Data Mining on SEER Data [14]

In this research article, the main focus is to develop a prediction model that can accurately predict survivability for lung cancer. SEER data is used for this project and WEKA tool for implementing the machine learning models at default parameters. Cancer survivability is done at multiple time periods like 6 months, 9 months, 1 year, 2 years and 5 years. Attribute selection techniques were also used to finally give out 13 attributes and also at the same time trying to retain the power of original attribute set. Machine learning models like support vector machines, neural nets, J48, random forest, logit boost, ensemble voting, and other classifiers were used. results have shown that voting ensemble have come out to be best predictor for lung cancer survivability. Accuracies of different survival time periods came out to be like 73.61% for 6 months, 74.45% for 9 months, 76.80% for 1 year, 85.45% for 2 years and finally 91.35% for 5 years were obtained from voting ensemble

classifier. Based on this study, an online calculator has been developed to predict the lung cancer outcome.

6. Reproducible Survival Prediction with SEER Cancer Data [15]

This is a review article in which it studies the previous research articles done on survivability of cancer using SEER data, to see if the studies have reproducibility results. They identified 34 articles and information is extracted about the experimental setup, it has been identified only one article has reproducible results and all the other studies have no direct reproducible results. The article also built machine learning models to predict the cancer survivability for 1 year and five years. They have used multilayer perceptron, logistic regression as the classification models. For the results, they have compared accuracy, area under the curve and f1 scores to determine the best performing model for prediction of cancer survivability. MLP has turned out to be the best predictor than other machine learning models.

7. Prediction of Breast Cancer Survivability Using Data Mining Techniques [16]

This article talks about prediction of breast cancer survivability using machine learning models, SEER data set is used for this project and after data preprocessing steps, a total of 16 attributed were selected for survival prediction analysis. In this research work, machine learning models like C4.5, Naïve Bayes and Artificial Neural Nets (back propagation) were used. This article clearly talks about creation of dependent variable and even compares with other research works. WEKA tool kit was used to build the machine learning models and also implemented 10-fold cross validation. C4.5 have turned out to be the best predictor of cancer survivability, multilayer perceptron came out to be second and naïve bayes has the least performance.

8. Stage Specific Predictive Models for Breast Cancer Survivability [17]

This research article focuses on a different approach for predicting the breast cancer survivability, they have gone through stage specific cancer to see if there is a difference in performance of the models trained across different stages. In this article, the authors have used three machine learning models like decision tree, logistic regression and naïve bayes. The models were trained on all stages all together and were also trained separately based on the stages. Based on the results, they concluded that the performance of the machine learning model is better when trained separately than when trained all together. The authors also found that the features which are important for predicting the survivability are different for different stages and the performance of the models also varied across stages.

9. Analysis of Cancer Data: A Data Mining Approach [18]

The main focus of this project is to build a machine learning model to predict the prostate cancer five-year survivability. Popular machine learning models like decision trees, support vector machines and artificial neural networks were used. A total of 77 attributes were used for cancer survivability, k-fold cross validation was used for evaluation purpose. The results have showed that support vector machines were best predictors for prostate cancer survival prediction with an accuracy of 92.85%, secondly is the neural nets and the least performing model is the decision tree.

10. Comparison of C4.5 and Naïve Bayes Classifier for the Prediction of Lung Cancer Survivability [19]

In this article, the main Focus is to create a machine learning model for prediction of lung cancer survivability and to see the effectiveness of the machine learning models from a historical data of 15years. For this project an open source tool called WEKA is used

for building the machine learning models like naive bayes and decision trees (J48) for predicting the survivability for five years, Seven years and ten years. Metrics like correctly classified instances, incorrectly classified instances, root mean squared error and kappa statistic were computed to determine the best performing model. The results show that J48 is better predictor than naïve bayes, it has a correctly classified instances of 94.44% for 5 years, 89.63% for 7 years and 90.61% for 10years.

11. Treating Colon Cancer Survivability Prediction as a Classification Problem [20]

This article talks about prediction of colon cancer survivability using the machine learning models, this study was performed using seer data set and after careful data preprocessing steps a total of 6 attributes were selected and prediction model was run using the data, later it was compared with the 18 attributes suggested by a physician. The main focus of this project is to see if machine learning models is as good as the model compared with the physicians. various machine learning techniques like Ada boost, KNN, naive bayes, decision tree, random forests were used in this study along with this, feature selection techniques and class balancing techniques were also implemented to improve the performance of the model. Finally, the results have shown that the performance of the model with six attributes is close to the performance of the model with 18 attributes suggested by the physician.

CHAPTER 3 : METHODS

3.1 DATA SOURCE

For this study, we used SEER (Surveillance Epidemiology and End Result) Cancer Incidence public database from the years 1973 – 2015. In order to obtain this data, we need to sign a Data Usage Agreement Form [21]. After approval, data can be downloaded from SEER cancer website using the credentials provided. The SEER program collects the incidence and survival data from the participating nine registries and provides the data to institutions for research purposes. SEER is a part of the Surveillance Research Program (SRP) under the National Cancer Institute (NCI). Currently SEER covers approximately 34.6% of the US population. The SEER program regularly collects data on patient demographics, primary tumor site, tumor morphology, stage at diagnosis, first course of treatment and vital status follow up. Mortality data is provided to SEER by the National Center for Health Statistics. It is said that the program is the only comprehensive source of population-based information that includes stages of cancer at diagnosis time and patient survival data in U.S. The SEER Cancer Incidence (1973-2015) public database consists of nine files which are in ASCII format. Each of these files contain data related to a particular anatomical site. The anatomical sites included are: (1) Breast Cancer, (2) Colon and Rectal Cancer, (3) Digestive and Thoracic Cancer, (4) Male Genital Cancer, (5) Female Genital

Cancer, (6) Lymphoma and Leukemia Cancer, (7) Urinary Cancer, (8) Respiratory Cancer, and (9) Other Cancers.

Each file contains 137 variables and each row in the file represents a cancer incidence. In total there are 5,094,238 records in all the cancer cohorts. The data obtained is the raw form and contains a lot of heterogeneous information like duplicate variables, deprecated variables, clinically irrelevant values and missing values, all these need to be dealt with in order to obtain processed data. Processed data then can be used to feed the machine learning models and predict cancer survivability. So, there is a need to understand the data and process it.

Out of the nine cancer cohorts present in the data file, we will not be considering two cancer cohorts. They are Lymphoma and Leukemia Cancers and Other Cancers. Lymphoma and Leukemia cancer cohort is not considered because, after performing all the data preprocessing steps, there were very few records to perform the survival prediction analysis. Other cancer cohort is not considered because, it contains a lot of granular information which can result in poor machine learning model performance.

3.2 DATA DICTIONARY

Data dictionary helps us understand the data, it gives information on every variable like the description of the variable, unique values present in the variable, time period, NAACCR names of the variables, how the variables were created and other important information. There is a detailed data dictionary file provided by SEER, it contains metadata and coding schema for all variables [22].

3.3 DATA PREPROCESSING

Data preprocessing is one of the most important steps in the field of machine learning since it takes a lot of time to understand the data and process it in the right methodology. Data preprocessing in this project is a multi-step process, during this we learn how to extract the data, convert the data into required formats, understand the independent variables which are important for the machine learning model, irrelevant records that needs to be eliminated, creating a dependent variable and various other steps which upon implementation produces a final output data that can be used for prediction analysis. For data preprocessing, we have used tools like KNIME and R programming.

KNIME is an open source and free platform for data analytics, machine learning and data mining. It integrates various components through its modular data pipelining concepts. A graphical user interface is present where we can drag and drop various types of nodes so that we can create a pipeline by connecting them together and execute various analytical processes, data mining and also later inspect the models. In this study, we used KNIME for variable selection, variable modification and to deal with missing values.

R programming language is also a free software environment for statistical computing and is used widely among statisticians, data miners etc. For this study, R programming is used to categorize the cohorts, creating the dependent variable and other necessary variables, perform basic data analysis, data mapping and creation of sparse format etc.

Figure 1: Data Preprocessing Steps

3.3.1 Data Extraction

Data obtained from SEER is in ASCII format and along with these ASCII files, SEER also provides a SAS code so that data can be converted into a readable form or converted to required formats. The ASCII format is converted into csv format.

3.3.2 Variable Selection

In the current study, each cancer cohort file obtained from SEER consists of 137 variables and apart from that we have created additional variables during the process. All the variables present in the dataset are not important for machine learning model prediction. In these, there are duplicate variables, deprecated variables and variables which are irrelevant for prediction or which do not have prediction capability. Feeding all these unnecessary variables to the machine learning model results in poor performance for prediction of five-year cancer survivability. So, in order to improve the performance, we need to select the right variables which have prediction capability. For this study we eliminated the variables by categorizing them into three types: (a) Elimination of

Deprecated Variables, (b) Elimination of Duplicate Variables, and (c) Elimination of Irrelevant Variables.

3.3.2.1 *Elimination of Deprecated Variables:*

In SEER data set, there are many variables which are deprecated i.e. there are variables which begin from considered time period but end abruptly in the middle, there are variables which begin in the middle of time period and continue until the end, there are variables which are present for a time period about 2 to 5 years, there are variables which are present for the whole time period, but the values in them are represented as NA's and are present only for some cancer cohorts, they are useful only for some particular cases. All these variables, even though they have prediction capability, they have a lot of missing values which impacts the machine learning model performance negatively, so they are eliminated. A total 74 variables were eliminated as deprecated variables (Table 1).

Table 1: Eliminated Deprecated Variables

S. No	Variables	NAACCR Name	Years active
1.	EOD10_SZ	EOD—Tumor Size	1988-2003
2.	EOD10_EX	EOD—Extension	1988-2003
3.	EOD10_PE	EOD—Extension Prost Path	1985-2003
4.	EOD10_ND	EOD—Lymph Node Involved	1988-2003
5.	EOD13	EOD—Old 13 Digit	1973-1982
6.	EOD2	EOD—Old 2 Digit	1973-1982
7.	EOD4	EOD—Old 4 Digit	1983-1987
8.	EOD_CODE	Coding System for EOD	1973-2003
9.	TUMOR_1V	Tumor Marker 1	1990-2003
10.	TUMOR_2V	Tumor Marker 2	1990-2003
11.	TUMOR_3V	Tumor Marker 3	1990-2003
12.	CSTUMSIZ	CS Tumor Size	2004+
13.	CSEXTEN	CS Extension	2004+

14.	CSLYMPHN	CS Lymph Nodes	2004+
15.	CSMETSDX	CS Mets at Dx	2004+
16.	CS1SITE	CS Site-Specific Factor 1	2004+
17.	CS2SITE	CS Site-Specific Factor 2	2004+
18.	CS3SITE	CS Site-Specific Factor 3	2004+
19.	CS4SITE	CS Site-Specific Factor 4	2004+
20.	CS5SITE	CS Site-Specific Factor 5	2004+
21.	CS6SITE	CS Site-Specific Factor 6	2004+
22.	CS25SITE	CS Site-Specific Factor 25	2004+
23.	DAJCCT	Derived AJCC T	2004+
24.	DAJCCN	Derived AJCC N	2004+
25.	DAJCCM	Derived AJCC M	2004+
26.	DAJCCSTG	Derived AJCC Stage Group	2004+
27.	DSS1977S	Derived SS1977	2004+
28.	DSS2000S	Derived SS2000	2004+
29.	DAJCCFL	Derived AJCC—Flag	2004+
30.	CSVFIRST	CS Version Input Original	2004+
31.	CSVLATES	CS Version Derived	2004+
32.	CSVCURRENT	CS Version Input Current	2004+
33.	SURGPRIF	RX Summ—Surg Prim Site	1998+
34.	SURGSCOF	RX Summ—Scope Reg LN Sur	2003+
35.	SURGSITF	RX Summ—Surg Oth Reg/Dis	2003+
36.	NUMNODES	RX Summ—Reg LN Examined	1998-2002
37.	SS_SURG	RX Summ—Surgery Type	1973-1997
38.	SURGSCOP	RX Summ—Scope Reg 98-02	1998-2002
39.	SURGSITE	RX Summ—Surg Oth 98-02	1998-2002
40.	AJCC_STG	AJCC stage 3rd edition (1988-2003)	1988-2003
41.	AJ_3SEER	SEER modified AJCC Stage 3rd ed (1988-2003)	1988-2003
42.	SSS77VZ	SEER Summary Stage 1977 (1995-2000)	1995-2000
43.	SSSM2KPZ	SEER Summary Stage 2000 (2001-2003)	2001-2003
44.	SUMM2K	Summary stage 2000 (1998+)	1998+
45.	CSTSEVAL	CS Tumor Size/Ext Eval	2004+
46.	CSRGEVAL	CS Lymph Nodes Eval	2004+

47.	CSMTEVAL	CS Mets Eval	2004+
48.	ER Status	ER Status Recode Breast Cancer (1990+)	1990+
49.	PR Status	PR Status Recode Breast Cancer (1990+)	1990+
50.	CS8SITE	CS Site-Specific Factor 8	2004+*
51.	CS10SITE	CS Site-Specific Factor 10	2004+*
52.	CS11SITE	CS Site-Specific Factor 11	2004+*
53.	CS13SITE	CS Site-Specific Factor 13	2004+*
54.	CS15SITE	CS Site-Specific Factor 15	2004+*
55.	CS16SITE	CS Site-Specific Factor 16	2004+*
56.	VASINV	Lymph vascular invasion	2004+*
57.	INSREC_PUB	Insurance recode (2007+)	2007+
58.	DAJCC7T	Derived AJCC-7 T	2010+
59.	DAJCC7N	Derived AJCC-7 N	2010+
60.	DAJCC7M	Derived AJCC-7 M	2010+
61.	DAJCC7STG	Derived AJCC-7 Stage Grp	2010+
62.	CS7SITE	CS Site-Specific Factor 7	2004+*
63.	CS9SITE	CS Site-Specific Factor 9	2004+*
64.	CS12SITE	CS Site-Specific Factor 12	2004+*
65.	HER2	Derived HER2 Recode (2010+)	2010+
66.	Brst_sub	Breast Subtype (2010+)	2010+
67.	CSMETSDB_PUB	CS Mets at Dx-Bone	2010+
68.	CSMETSDBR_PUB	CS Mets at Dx-Brain	2010+
69.	CSMETSDDLIV_PUB	CS Mets at Dx-Liver	2010+
70.	CSMETSDDLUNG_PUB	CS Mets at Dx-Lung	2010+
71.	T_VALUE	T value - based on AJCC 3rd (1988-2003)	1988-2003
72.	N_VALUE	N value - based on AJCC 3rd (1988-2003)	1988-2003
73.	M_VALUE	M value - based on AJCC 3 rd (1988-2003)	1988-2003
74.	RAD_BRNR	Radiation to Brain or CNS Recode (1988-1997)	1988-1997

3.3.2.2 Elimination of Duplicate Variables

In the SEER cancer dataset, there are duplicate variables which hinder the performance of machine learning models. A total of 25 variables were identified and eliminated. As we are considering staging variable, we are eliminating T, N, M staging variables under this variable category. The duplicate variables which are eliminated mentioned below (Table 2).

Table 2: Eliminated Duplicate Variables

S. No	Variables	NAACCR Name
1.	HISTO2V	Histology (92-00) ICD-O-2
2.	BEHO2V	Behavior (92-00) ICD-O-2
3.	SITERWHO	Site Recode ICD-O-3/WHO 2008
4.	ICDOT09V	Recode ICD-O-2 to 9
5.	ICDOT10V	Recode ICD-O-2 to 10
6.	ICCC3WHO	ICCC site recode ICD-O-3/WHO 2008
7.	ICCC3XWHO	ICCC site rec extended ICD-O-3/WHO 2008
8.	BEHTREND	Behavior Recode for Analysis
9.	HISTREC	Histology Recode—Broad Groupings
10.	CS0204schema	CS Schema v0204+
11.	RAC_RECA	Race recode (White, Black, Other)
12.	RAC_RECY	Race recode (W, B, AI, API)
13.	ORIGRECB	Origin recode NHIA (Hispanic, NonHisp)
14.	HST_STGA	SEER historic stage A
15.	CODPUB	Cause of Death to SEER site recode
16.	CODPUBKM	COD to site rec KM
17.	AYASITERWHO	AYA site recode/WHO 2008
18.	LYMSUBRWHO	Lymphoma subtype recode/WHO 2008
19.	Intprim	Primary by international rules
20.	CSschema	CS Schema -AJCC 6 th ed (previously called v1)
21.	ADJAJCCSTG	Breast Adjusted AJCC 6 th Stage (1988+)
22.	ANNARBOR	Lymphomas: Ann Arbor Staging (1983+)
23.	ADJTM_6VALUE	Breast Adjusted AJCC 6 th T (1988+)
24.	ADJNM_6VALUE	Breast Adjusted AJCC 6 th N (1988+)
25.	ADJM_6VALUE	Breast Adjusted AJCC 6 th M (1988+)

3.3.2.3 Elimination of Irrelevant Variables

In the current dataset there are variables which are not related to the current prediction model. They consume a lot of time and memory; those variables are eliminated.

Table 3 shows the list of irrelevant variables which are eliminated.

Table 3: Eliminated Irrelevant Variables

S. No	Variables	NAACCR Name
1.	PUBCSNUM	Patient ID number
2.	REG	Registry ID
3.	YR_BRTH	Year of Birth
4.	MDXRECOMP	Month of diagnosis
5.	YEAR_DX	Year of diagnosis
6.	DX_CONF	Diagnostic Confirmation
7.	REPT_SRC	Type of Reporting Source
8.	EOD10_NE	Regional Nodes Examined
9.	REC_NO	SEER Record Number
10.	COHORT	Type of cancer
11.	srv_time_mon	Survival months
12.	srv_time_mon_flag	Survival months flag
13.	VSRTSADX	SEER Cause-Specific Death Classification
14.	ODTHCLASS	SEER Other Cause of Death Classification
15.	ST_CNTY	State-county recode
16.	CODPUB	Cause of Death to SEER site recode
17.	CODPUBKM	COD to site rec KM
18.	STAT_REC	Vital Status recode
19.	IHSLINK	his Link
20.	TYPE_FU	SEER Type of Follow-up
21.	UNIQ_ID	Unique number to each record

After eliminating all these variables, we are left with a total of 22 variables, of which 21 are called independent variables and the other is dependent variable. These

independent variables are used in the prediction for cancer survivability and the dependent variable is the class column or target variable. Table 4 lists the variables which are used for survival prediction in this project.

Table 4: Variables Chosen for Prediction of Cancer Survivability

S. No	Variables	NAACCR Name
1.	MAR_STAT	Marital Status at DX
2.	RACE1V	Race/Ethnicity
3.	NHIADE	NHIA Derived Hispanic Origin
4.	SEX	Sex
5.	AGE_DX	Age at diagnosis
6.	SEQ_NUM	Sequence Number—Central
7.	PRIMSITE	Primary Site
8.	LATERAL	Laterality
9.	HISTO3V	Histologic Type ICD-O-3
10.	BEHO3V	Behavior Code ICD-O-3
11.	GRADE	Grade
12.	EOD10_PN	Regional Nodes Positive
13.	NO_SURG	Reason for no surgery
14.	FIRSTPRM	First malignant primary indicator
15.	MALIGCOUNT	Total Number of In Situ/malignant Tumors for Patient
16.	BENBORDCOUNT	Total Number of Benign/Borderline Tumors for Patient
17.	RADIATNR	Radiation Recode
18.	RAD_SURG	Radiation Sequence with Surgery
19.	CHEMO_RX_REC	Chemotherapy recode (yes, no/unk)
20.	TUMOR SIZE	
21.	CANCER_STAGING	
22.	FIVE_YEAR_SURVIVAL	

In the afore mentioned tables, there are certain variables like tumor size, cancer staging, five-year survival, unique ID and cohort. All these variables were created during data preprocessing steps and will be discussed shortly.

Figure 2: Variable Selection Process

3.3.3 Merging of Identical Variables

All the deprecated variables which are eliminated are not useless, there are some variables which have important information about cancer. After careful observation, we came to a conclusion to merge tumor size variables, there are multiple variables which give information about tumor size and they are present only for a short period of time. The tumor size variables are merged since their schema is similar and there is no need to modify any unique values. EOD Tumor Size variable gives information on size of tumor from 1988 to 2003 and from there the variable is deprecated. The schema of the variable is simple, if the code given is 1, then the tumor size is 1 millimeter. If the code given is 10, then the size of tumor is 10mm which is 1cm. A similar type of schema is present for CS Tumor Size which started from the year 2004 and continued until the end of the time period. Figure 3 provides detailed information on parent variable and merged variable.

Figure 3: Merging of Tumor Size Variable

3.3.4 Data Mapping

There are multiple staging variables which are categorized as deprecated at the beginning, but they contain useful information regarding cancer and also have a very important role in prediction for cancer survivability. These variables are present in two different time periods and in this case, they do not have same coding schema. These two variables are coded into a new variable called cancer staging. The two staging variables are: (a) AJCC Stage 3 (1988 – 2003), and (b) Derived AJCC Stage 6 (2004+). These two variables were recoded into five categories as Stage 0, Stage I, Stage II, Stage III and Stage IV, they were given unique codes like 0, 1, 2, 3, 4.

3.3.4.1 AJCC Stage 3 (1988 – 2003)

This variable gives information about stage of cancer from the year 1988 to 2003, it is derived by an algorithm from Extend of Disease (EOD). Below we will be discussing how AJCC Stage 3 variable has been recoded into Cancer Staging variable.

1. In this variable code 0 gives information about Insitu cancer type, which is also known as Stage 0, this unique value has been recoded to 0 in Cancer Staging variable, which represents Stage 0.
2. From code 10 to 19 gives information about Stage I and its subtypes, all these unique values were recoded to unique value 1 in Cancer Staging variable, which represents Stage I.
3. From code 20 to 29 gives information about Stage II and its subtypes, all these values were recoded to unique value 2 in Cancer Staging variable, which represents Stage II.
4. From code 30 to 39 gives information about Stage III and its subtypes, all these values were recoded to unique value 3 in Cancer Staging variable, which represents Stage III.
5. From code 40 to 49 gives information about Stage IV and its subtypes, all these values were recoded to unique value 4 in Cancer Staging variable, which represents Stage IV.
6. Remaining unique values represents N/A's, error condition etc. which were not coded and left as they were.

3.3.4.2 Derived AJCC Stage 6 (2004+):

This variable gives information about stages of cancer from the year 2004 till the end of considered time period. For this variable, codes were derived from CS detailed site-specific codes using CS algorithm. Below we will be discussing on how Derived AJCC Stage 6 variable has been recoded into Cancer Staging variable.

1. In this variable from code 00 to 02 gives information about Stage 0, these unique values have been recoded to unique value 0 in Cancer Staging variable, which represents Stage 0.
2. From code 10 to 24 gives information about Stage I and its subtypes, all these unique values were recoded to unique value 1 in Cancer Staging variable, which represents Stage I.
3. From code 30 to 43 gives information about Stage II and its subtypes, all these values were recoded to unique value 2 in Cancer Staging variable, which represents Stage II.
4. From code 50 to 63 gives information about Stage III and its subtypes, all these values were recoded to unique value 3 in Cancer Staging variable, which represents Stage III.
5. From code 70 to 74 gives information about Stage IV and its subtypes, all these values were recoded to unique value 4 in Cancer Staging variable, which represents Stage IV.
6. Remaining unique values represents N/A's, stage unknown etc. which were not coded and left as they were.

After recoding the variables AJCC Stage 3 (1988 – 2003) and Derived AJCC Stage 6 (2004+) into a new variable called Cancer Staging, which gives information about stages of cancer from years 1988 to 2015. The new variable has the coding schema as mentioned below in Table 5.

Table 5: Coding Schema for Cancer Staging Variable

Cancer Staging	
Code	Stage of Cancer
0	Stage 0
1	Stage I
2	Stage II
3	Stage III
4	Stage IV

3.3.5 Dependent Variable Creation

After performing various steps like Variable Selection, Variable Merging and Data Mapping, we are finally left with 21 variables which are called as features or Independent variables. These independent variables help in predicting five-year cancer survival accurately. In order to predict the outcome, we need a target variable/dependent variable. In the current study, we created a binary variable which represents whether the person has survived or not survived. This variable is created based on specific conditions by considering the variables like Type of Follow up, Specific Cause of Death, Survived Months, and Vital Status. Below we will be discussing the variables used to create dependent variable and the conditions on which the dependent variable is created.

Type of Follow up: This variable gives information on whether the case is an active follow up or not. Active follow up means, the person is regularly checking up with the medical staff to update about the condition.

Specific Cause of Death: The variable gives information about whether the patient is dead with the condition he was diagnosed with or not.

Survived Months: This variable gives information on the number of months the relative case has survived after the diagnosis.

Vital Status: This variable gives information whether the specified case is currently alive or dead.

Based on the above-mentioned variables, we created the dependent variable based on conditions like:

1. Cases with active follow up are considered for creation of Dependent variable.
2. Cases with Survived Months less than or equal to 60 and Vital Status is 'dead' and Specific Cause of Death is 'Yes' then those records are considered as "Not Survived".
3. Cases with Survived Months equal to 60 and vital status recode is alive then those records were considered as "Survived".
4. Cases with Survived Months greater than 60 were considered "Survived" no matter the cause of death, because the patient has survived the given amount of time.

All the remaining records were considered not applicable and later eliminated.

Figure 4: Flow Diagram of Creation of Dependent Variable

Type of records which are not considered for this project are:

1. Records which have Vital Status as Alive and Survived Months less than 60.
2. Records with no active follow up.
3. Records with Survived Months less than 60, Vital Status is Dead but Cause Specific Death is NO, those records were not considered for the model prediction.

The binary variable created is called “FIVE-YEAR SURVIVAL” and the unique values created as (a) Y – Survived cases, and (b) N - Not Survived cases.

3.3.6 Record Elimination with Clinically Irrelevant Values

Every variable has some values like 9, 99, 999 and these values represent unknown values or N/A's. These unique values will affect the performance of the machine learning models. These values are considered as noise in the data and to be removed to improve the model performance. Below we will be discussing on how clinically irrelevant values were eliminated in a cohort using Breast cancer as an example.

There are 828,457 records in the cohort and after eliminating the records which are not considered while creating the dependent variable, there are 604,895 records. Then we eliminated the unknown value which is coded as '9' in Marital Status and the number of records came down to 583,130. Then we eliminated the clinically irrelevant value which is coded as '99' in Race and then number of records came down to 582,185. Next we eliminated the irrelevant values in Age variable, for this study we chose age range from 18 years to 110 years and remaining records were eliminated. Then we are left with 582,156 records. Next we chose to eliminate the irrelevant values in SEQUENCE NUMBER, values greater than 87 were eliminated since they fall under the category of Unspecified or Unknown, this left us with 582,146 records. Then we chose to eliminate the irrelevant values from variable GRADE, this resulted in elimination of more than 200,000 records, so we chose not to eliminate these values in case of this variable. then we chose Positive Nodes to eliminate the irrelevant value which is coded as '99' and this has left us with 433,798 records. Then we eliminated the irrelevant value which is coded as '9' from

Reason for No Surgery variable. It describes that Unknown if surgery is performed: Autopsy or Death Certificate. After eliminating that value, we are left with 433,690 records. Then we eliminated records from Cancer Staging with irrelevant values and this has left with 405,599 records. Then finally we eliminated irrelevant records from Tumor Size, in Delens 2004 article, it mentioned that if the tumor size is >200 mm then it might be error in data entry point. So, we chose to eliminate other values, and this has left us finally with 367,403 records. After applying all these steps, the data set is free from all clinically irrelevant values.

Eliminating all these clinically irrelevant values and processing the data, i.e. the values which are chosen in each and every variable can be considered as Eligibility Criteria for the current study. Similar eligibility criteria applied to remaining cancer cohorts.

3.3.7 Record Elimination with Missing Values

Missing values in a dataset affects the performance of model prediction, so we need to deal with the missing data. There are multiple ways to deal with the missing data like (a) Use Most Frequent Values, (b) Use Mean/Median Values, (c) Use Previous or Next Values, and (d) Eliminate Records with Missing Values

For this study, we chose to eliminate the records with missing values because [23], we do not want to change the testing data which might have impact on the outcome. But previous data preprocessing steps have eliminated the records with missing values from all the cancer cohorts.

After all these data preprocessing steps we are finally left with the number of records for each cohort mentioned in the Table 6.

Table 6: Cohort Population

COHORT	No of Patients in SEER Data Set	No of Patients Included in Study After Applying Eligibility Criteria
Breast	828,457	367,403
Colon & Rectal	554,686	165,168
Digestive & Thoracic	393,963	47,946
Male Genital	673,417	40,400
Female Genital	333,245	57,772
Respiratory	682,112	167,742
Urinary	345,897	53,348

All the data preprocessing steps like Creation of Cohort Variable, Creation of Dependent Variable, Creation of Cancer Staging Variable are implemented in R program, remaining steps like Variable Selection, Variable Merging, Record Elimination with Clinically Irrelevant Values and Record Elimination with Missing Values are implemented in KNIME.

3.3.8 Creation of Sparse Format:

After removing the missing values, the dataset is loaded into a machine learning model in python and it prompted an error saying that it cannot accept characters to run the model. There are two variables which are present in the data set with characters, PRIMESITE and FIVE-YEAR SURVIVAL, so we converted the data into sparse format. In order to create the sparse format, the variable should contain values other than numbers. Creating the sparse format causes an increase in the number of variables, this method converts unique values present in a variable into each individual column and fills them with binary values either 0 or 1 (Figure 5).

S. No	FIVE-YEAR SURVIVAL
1.	Y
2.	N
3.	Y
4.	Y
5.	N

S. NO	FIVE-YEAR SURVIVAL_Y	FIVE-YEAR SURVIVAL_N
1.	1	0
2.	0	1
3.	1	0
4.	1	0
5.	0	1

Figure 5: Creation of Sparse Format

Creation of Sparse Format is done using R program and after performing the sparse format, we need to remove the parent variables (PRIMESITE, FIVE-YEAR SURVIVAL) and FIVE-YEAR SURVIVAL_N to not affect the prediction model performance.

3.4 EXPERIMENTAL APPROACH

After implementing all the data preprocessing steps, we now can feed the data into machine learning models and predict the five-year survivability. For this study we implemented 8 types of machine learning models.

1. Decision Tree (C4.5) – **DT1**
2. Decision Tree (Gini Index) – **DT2**
3. Logistic Regression - **LR**
4. Naïve Bayes - **NB**
5. Random Forest (Info Gain) – **RF1**
6. Random Forest (Gini Index) – **RF2**
7. Support Vector Machine - **SVM**
8. Multi-Layer Perceptron - **MLP**

We also implemented class balancing techniques which to a total gives us additional cohorts to perform prediction analysis. The class balancing techniques are:

1. No Sampling
2. Over Sampling
3. Under Sampling

We will be discussing the implemented class balancing techniques and their methods in the next chapters. Out of nine cohorts, we used seven cohorts in this study, they are: (1) Breast Cancer, (2) Colon & Rectal Cancer, (3) Digestive & Thoracic Cancer, (4) Female Genital Cancer, (5) Male Genital Cancer, (6) Respiratory Cancer, and (7) Urinary Cancer.

Based on all the machine learning models, sampling techniques and cancer cohorts, we designed the experimental approach in a way that every machine learning algorithm runs with all the types of sampling techniques in every cohort. This finally yields a total of 168 experiments which are compared together to find out the best machine learning model for survival prediction and also to conclude whether the sampling techniques have improved the performance of machine learning models.

Figure 6: Machine Learning Experimental Approach

3.5 CROSS-VALIDATION

Cross-Validation is a technique used in machine learning to remove the bias between training data and testing data. There are multiple cross-validation techniques like (a) Holdout Method, (b) K-Fold Cross-Validation, (c) Stratified K-Fold Cross-Validation, and (d) Leave-P-Out Cross-Validation.

In the current study we implemented Stratified 10-Fold Cross-Validation [10] . In this method, the dataset is divided into 10 folds where each fold contains the same percentage of samples of each target class. By performing this technique, we can minimize the bias between training and testing datasets. In each fold nine datasets are categorized as training dataset and the machine learning model trains on it and the model is run against the 10th subset. This process repeats on every fold and the results were obtained as the average of all the folds.

Equation 1: Cross Validation Accuracy (CVA)

$$CVA = \frac{1}{k} \sum_{i=1}^k A_i$$

Fold #	Dataset split into K (10) parts									
1	Blue	Green	Green	Green	Green	Green	Green	Green	Green	Green
2	Green	Blue	Green	Green	Green	Green	Green	Green	Green	Green
3	Green	Green	Blue	Green	Green	Green	Green	Green	Green	Green
4	Green	Green	Green	Blue	Green	Green	Green	Green	Green	Green
5	Green	Green	Green	Green	Blue	Green	Green	Green	Green	Green
6	Green	Green	Green	Green	Green	Blue	Green	Green	Green	Green
7	Green	Green	Green	Green	Green	Green	Blue	Green	Green	Green
8	Green	Green	Green	Green	Green	Green	Green	Blue	Green	Green
9	Green	Green	Green	Green	Green	Green	Green	Green	Blue	Green
10	Green	Green	Green	Green	Green	Green	Green	Green	Green	Blue

Figure 7: Cross Validation (Green Color: Training Data, Blue Color: Testing Data)

3.6 METRICS

3.6.1 Sensitivity

Sensitivity is the measure of proportion of positive cases that actually predicted as positive. It is the ratio of True Positive to sum of True Positive and False Negative. It can also be called as recall.

$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

3.6.2 Accuracy

Accuracy is the measure of number of correct predictions to the total number of input samples. It is the ratio of sum of True Positive and True Negative to sum of True Positive, True Negative, False Positive and False Negative.

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

3.6.3 Specificity

Specificity is the measure of proportion of negative cases that actually predicted as negative. It is the ratio of True Negative to the sum of True Negative and False Positive.

$$\text{Specificity} = \frac{TN}{TN+FP}$$

3.6.4 Precision

Precision is the ratio of True Positive to the sum of True Positive and False Positive.

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

3.6.5 F1 Score

F1 score is defined as weighted harmonic mean of tests precision and recall.

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

Or

$$\text{F1 Score} = \frac{2TP}{2TP + FP + FN}$$

3.6.6 AUC Score

AUC score is used to determine the performance of the machine learning models. Higher the value of AUC, better is the model. This metric is useful while comparing multiple classification models.

All the metrics were calculated for each machine learning model and they were compared with each other across multiple sampling techniques to find out the better sampling technique and across the cohorts to find out the better performing model for five-year cancer survivability.

CHAPTER 4 : CLASS BALANCING TECHNIQUES

Datasets obtained after preprocessing were highly imbalanced across all cancer cohorts (Table 7). Machine learning models tend to work better in cases of balanced datasets. Models built on imbalanced datasets tend to lean towards majority class thereby creating bias, which in turn leads to higher misclassification of minority class and poor performance of the models. Balancing the classes improve the prediction performance of the models and reduce the chances of misclassification. Class balancing techniques when applied to whole datasets which is a flaw, this leads to over optimistic or over-fitted models [24]. In the current project, the class balancing technique was implemented only on the training dataset, but not on testing dataset. This prevents overfitting of the models [25].

Table 7: Imbalanced Cohorts

Cohort	Total Cancer Cases (records) included in the ML Analysis	# of Record (5-Year Survival=Y)	# of Record (5-Year Survival=N)
Breast	367,403	333,559 (90.79%)	33,844 (9.21%)
Colon & Rectum	165,168	105,736 (64.02%)	59,432 (35.98%)
Digestive & Thoracic	47,946	7,739 (16.14%)	40,207 (83.86%)
Male Genital	40,400	38,902 (96.29%)	1,498 (3.71%)
Female Genital	57,772	42,624 (73.78%)	15,148 (26.22%)
Respiratory	167,742	36,706 (21.88%)	131,036 (78.12%)
Urinary	53,348	39,757 (74.52%)	13,591 (25.48%)

There are many ways to deal with imbalanced data sets, some of them are: (1) Over Sampling, (2) Under Sampling, and (3) Generating synthetic data. For this project we have chosen to perform sampling techniques like oversampling and undersampling.

4.1 OVER SAMPLING

Over Sampling increases the instances of minority class to that of majority class till the distribution balance is attained. There are many methods in over sampling technique, like Random Over Sampling, SMOTE etc. In the current study we chose Random Over Sampling (ROS) technique. It involves random repetition of minority class instead of repeating every sample to make it equal to that of majority class. ROS has proven to be a robust technique in case of Over Sampling [24]. Balanced Over Sampling datasets for the cancer cohorts can be seen in Table 8.

Table 8: Random Over Sampling of Training Data.

Cohort	Total Cancer Cases (records) included in the ML Analysis	# of Over Sampled Records (5-Year Survival=Y)	# of Over Sampled Records (5-Year Survival=N)
Breast	367,403	300,204 (50%)	300,204 (50%)
Colon & Rectum	165,168	95,163 (50%)	95,163 (50%)
Digestive & Thoracic	47,946	36,187 (50%)	36,187 (50%)
Male Genital	40,400	35,012 (50%)	35,012 (50%)
Female Genital	57,772	38,362 (50%)	38,362 (50%)
Respiratory	167,742	117,932 (50%)	117,932 (50%)
Urinary	53,348	35,783 (50%)	35,782 (50%)

4.2 UNDER SAMPLING

Under Sampling technique eliminates the records from the majority class and make the number of instances equal to that of minority class till the distribution balance is attained. There are multiple methods used in case of undersampling, like Random Under Sampling, Cluster centroids etc. For this study, we chose Random Under Sampling technique, it involves elimination of records randomly from majority class and makes the instances equal to that of minority class. Under Sampling techniques were not used quite

often because there is possible loss of useful information [24,25]. Balanced Under Sampling data sets for multiple cancer cohorts can be seen in Table 9.

Table 9: Random Under Sampling of Training Data.

Cohort	Total Cancer Cases (records) included in the ML Analysis	# of Under Sampled Records (5-Year Survival=Y)	# of Under Sampled Records (5-Year Survival=N)
Breast	367,403	30,459 (50%)	30,459 (50%)
Colon & Rectum	165,168	53,489 (50%)	53,489 (50%)
Digestive & Thoracic	47,946	6,965 (50%)	6,965 (50%)
Male Genital	40,400	1,348 (50%)	1,348 (50%)
Female Genital	57,772	13,633 (50%)	13,633 (50%)
Respiratory	167,742	33,036 (50%)	33,036 (50%)
Urinary	53,348	12,232 (50%)	12,232 (50%)

Imbalanced data class distribution has a major impact on classifying the instances correctly. Common experimental flaw was implementing class balancing techniques to the whole dataset. This leads to overfitting of the model and over optimistic values even though the Cross Validation was implemented properly [26]. It means the machine learning model is predicting values on the data it has already been trained on [24]. In the current project, class balancing techniques were implemented only on the training datasets at each fold and testing datasets were left as they are to prevent model overfitting. Class balancing techniques not only balances the major and minor class instances in the dataset, but it also improves the performance of the machine learning model.

CHAPTER 5 : RESULTS

The results obtained in this study by implementing 10-fold Cross Validation were compared across multiple machine learning models and also across multiple sampling techniques which are implemented. Obtained results were in the form of a confusion matrix from which the metrics were calculated. The confusion matrix presents values belonging to 4 categories, such as True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). True Positive and True Negative denotes the number of records that were classified correctly. False Positive denotes the number of cases that are predicted as positive but were actually negative. False Negative denotes the number of cases classified as negative but were actually positive. The performance of the models evaluated by calculating the metrics like Sensitivity, Accuracy, Specificity, Precision and F1 score and comparing the obtained values across machine learning models and across sampling techniques. AUC values were also computed to give an idea on the performance of the model.

5.1 BREAST CANCER

In case of Breast Cancer, from the results it's evident that the specificity is poor in case of no sampling technique and it has improved upon implementing oversampling and undersampling techniques. When there was no sampling technique implemented (Table 10), Naive Bayes has performed better even though sensitivity and accuracy are too low,

but specificity of the model was higher compared to other models. Decision Trees came second, then Random Forests, Logistic Regression, Multi-Layer Perceptron, and the least performing model was Support Vector Machine. For oversampling and undersampling techniques (Table 11 & Table 12), Random Forests have outperformed other models, then came Support Vector Machines, and Naive Bayes has least performance.

Table 10: Breast Cancer No Sampling Results

Algorithms	Sensitivity	Accuracy	Specificity	Precision	F1-score
Decision Tree (C4.5)	0.9709	0.9194	0.4119	0.9421	0.9563
Decision Tree (Gini)	0.9714	0.9198	0.411	0.942	0.9565
Logistic Regression	0.9883	0.9268	0.3205	0.9348	0.9608
Naïve Bayes	0.7665	0.7662	0.764	0.9697	0.8562
Random Forest (Info)	0.9908	0.9294	0.3244	0.9353	0.9623
Random Forest (Gini)	0.9903	0.9295	0.3308	0.9358	0.9623
Support Vector Machine	0.9932	0.9256	0.259	0.9296	0.9604
Multi-Layer Perceptron	0.9863	0.9256	0.3278	0.9353	0.9601

Table 11: Breast Cancer Over-Sampling Results

Algorithms	Sensitivity	Accuracy	Specificity	Precision	F1-score
Decision Tree (C4.5)	0.859	0.8379	0.6298	0.9581	0.9058
Decision Tree (Gini)	0.8517	0.8327	0.645	0.9594	0.9024
Logistic Regression	0.783	0.7845	0.7998	0.9747	0.8684
Naïve Bayes	0.536	0.569	0.8943	0.9804	0.6931
Random Forest (Info)	0.8146	0.8152	0.8206	0.9781	0.8889
Random Forest (Gini)	0.8217	0.8208	0.8125	0.9774	0.8928
Support Vector Machine	0.7907	0.7911	0.7946	0.9743	0.873
Multi-Layer Perceptron	0.798	0.7968	0.7848	0.9734	0.877

Table 12: Breast Cancer Under Sampling Results

Algorithms	Sensitivity	Accuracy	Specificity	Precision	F1-score
Decision Tree (C4.5)	0.7749	0.7774	0.8017	0.9747	0.8634
Decision Tree (Gini)	0.7758	0.7783	0.8022	0.9748	0.864
Logistic Regression	0.784	0.7852	0.7964	0.9743	0.8689
Naïve Bayes	0.5404	0.5729	0.8934	0.9804	0.6968
Random Forest (Info)	0.8112	0.8124	0.8247	0.9785	0.887
Random Forest (Gini)	0.8086	0.8102	0.8264	0.9787	0.8855
Support Vector Machine	0.7915	0.7918	0.7939	0.9743	0.8734
Multi-Layer Perceptron	0.7662	0.7678	0.7837	0.9721	0.857

5.2 COLON & RECTAL CANCER

For Colon & Rectal Cancer, implementing the sampling techniques have improved the performance of the machine learning models. From Table 13, we can say that Random Forests have outperformed the other models, Logistic Regression came second, then Support Vector Machines, Decision Trees, Multi-Layer Perceptron and finally Naive Bayes with least performance when no sampling techniques were implemented. For oversampling and undersampling techniques (Table 14 & Table 15), it's clear that Random Forests have outperformed all other machine learning models. Support Vector Machine

and Logistic Regression came as second-best predictors, then Decision Trees, and Multi-Layer Perceptron. Naive Bayes came out as the least predictor of all models.

Table 13: Colon & Rectal Cancer No Sampling Results

Algorithms	Sensitivity	Accuracy	Specificity	Precision	F1-score
Decision Tree (C4.5)	0.8616	0.7955	0.6778	0.8263	0.8436
Decision Tree (Gini)	0.8599	0.7971	0.6854	0.8294	0.8444
Logistic Regression	0.8736	0.8056	0.6845	0.8313	0.8519
Naïve Bayes	0.4983	0.6424	0.8987	0.8975	0.6408
Random Forest (Info)	0.923	0.8243	0.6488	0.8238	0.8706
Random Forest (Gini)	0.9219	0.8242	0.6505	0.8243	0.8704
Support Vector Machine	0.8783	0.8075	0.6815	0.8307	0.8539
Multi-Layer Perceptron	0.8715	0.7705	0.5907	0.7911	0.8294

Table 14: Colon & Rectal Cancer Over Sampling Results

Algorithms	Sensitivity	Accuracy	Specificity	Precision	F1-score
Decision Tree (C4.5)	0.7833	0.7679	0.7403	0.8429	0.812
Decision Tree (Gini)	0.7839	0.7701	0.7454	0.8456	0.8136
Logistic Regression	0.7833	0.7838	0.7847	0.8662	0.8227
Naïve Bayes	0.4567	0.6192	0.9083	0.8986	0.6057
Random Forest (Info)	0.8382	0.8116	0.7643	0.8635	0.8507
Random Forest (Gini)	0.8368	0.8109	0.7648	0.8636	0.85
Support Vector Machine	0.7829	0.7856	0.7903	0.8691	0.8238
Multi-Layer Perceptron	0.7642	0.7702	0.7808	0.8612	0.8098

Table 15: Colon & Rectal Cancer Under Sampling Results

Algorithms	Sensitivity	Accuracy	Specificity	Precision	F1-score
Decision Tree (C4.5)	0.7695	0.7705	0.7722	0.8574	0.8111
Decision Tree (Gini)	0.7711	0.7716	0.7727	0.8578	0.8121
Logistic Regression	0.784	0.7852	0.7964	0.9743	0.8689
Naïve Bayes	0.4508	0.6158	0.9094	0.8985	0.6004
Random Forest (Info)	0.8353	0.8109	0.7675	0.8647	0.8498
Random Forest (Gini)	0.8322	0.8098	0.77	0.8655	0.8485
Support Vector Machine	0.7827	0.7856	0.7907	0.8693	0.8238
Multi-Layer Perceptron	0.6562	0.7088	0.8022	0.8551	0.7426

5.3 DIGESTIVE & THORACIC CANCER

The results of Digestive & Thoracic Cancer indicated that sensitivity of the models in no sampling technique were poor and implementing sampling techniques have improved the sensitivity of the models. For no sampling technique (Table 16), Random Forests have outperformed all other machine learning models. Decision Trees came second, then Logistic Regression, Support Vector Machines, Multi-Layer Perceptron. Naive Bayes turned out to be the least predictor of all models. When sampling techniques were implemented (Table 17 & Table 18), Random Forests have again outperformed other machine learning models. Support Vector Machines came out as second best predictors, then Decision Trees, Logistic Regression, Multi-Layer Perceptron. Naive Bayes as usual turned out to be with least performance.

Table 16: Digestive & Thoracic Cancer No Sampling Results

Algorithms	Sensitivity	Accuracy	Specificity	Precision	F1-score
Decision Tree (C4.5)	0.5714	0.8898	0.9511	0.6923	0.6261
Decision Tree (Gini)	0.5705	0.8886	0.9498	0.6864	0.6231
Logistic Regression	0.499	0.8934	0.9693	0.7576	0.6017
Naïve Bayes	0.5767	0.8333	0.8827	0.4863	0.5276
Random Forest (Info)	0.5302	0.9035	0.9754	0.8058	0.6395
Random Forest (Gini)	0.5542	0.906	0.9737	0.802	0.6555
Support Vector Machine	0.4193	0.89	0.9806	0.8058	0.5516
Multi-Layer Perceptron	0.4437	0.8875	0.973	0.7596	0.5602

Table 17: Digestive & Thoracic Cancer Over Sampling Results

Algorithms	Sensitivity	Accuracy	Specificity	Precision	F1-score
Decision Tree (C4.5)	0.7255	0.8528	0.8773	0.5323	0.6141
Decision Tree (Gini)	0.7285	0.8495	0.8727	0.5242	0.6097
Logistic Regression	0.8341	0.8122	0.808	0.4554	0.5892
Naïve Bayes	0.5837	0.8295	0.8768	0.477	0.525
Random Forest (Info)	0.8537	0.8575	0.8583	0.5369	0.6592
Random Forest (Gini)	0.8312	0.8708	0.8785	0.5683	0.6751
Support Vector Machine	0.8465	0.8242	0.8199	0.475	0.6085
Multi-Layer Perceptron	0.7647	0.7944	0.8001	0.424	0.5455

Table 18: Digestive & Thoracic Cancer Under Sampling Results

Algorithms	Sensitivity	Accuracy	Specificity	Precision	F1-score
Decision Tree (C4.5)	0.8306	0.8277	0.8271	0.4805	0.6088
Decision Tree (Gini)	0.8287	0.8284	0.8284	0.4811	0.6088
Logistic Regression	0.839	0.8148	0.8101	0.4596	0.5939
Naïve Bayes	0.5679	0.8323	0.8832	0.4835	0.5223
Random Forest (Info)	0.8854	0.834	0.8241	0.4921	0.6326
Random Forest (Gini)	0.8775	0.84	0.8327	0.5024	0.639
Support Vector Machine	0.8484	0.8234	0.8186	0.4738	0.608
Multi-Layer Perceptron	0.8522	0.6789	0.6456	0.3164	0.4614

5.4 MALE GENITAL CANCER

In case of Male Genital Cancer, the specificity of the models in no sampling technique were poor and with oversampling and undersampling techniques implemented, the specificity has improved significantly across all models. Random Forests and Decision Trees were on par with each other, Logistic Regression came second, then Multi-Layer Perceptron, and Naive Bayes. Support Vector Machines turned out to be the least predictor in no sampling technique (Table 19). For oversampling & undersampling techniques (Table 20 & Table 21), Random Forests have outperformed all other machine learning models. Support Vector Machines came out to be second best predictors, then Logistic Regression, Decision Trees and Multi-Layer Perceptron. Naive Bayes turned out to be with least performance of all.

Table 19: Male Genital Cancer No Sampling Results

Algorithms	Sensitivity	Accuracy	Specificity	Precision	F1-score
Decision Tree (C4.5)	0.9916	0.9666	0.3158	0.9741	0.9828
Decision Tree (Gini)	0.9921	0.9672	0.3224	0.9744	0.9831
Logistic Regression	0.9972	0.9678	0.2036	0.9702	0.9835
Naïve Bayes	0.9121	0.8903	0.3244	0.9723	0.9412
Random Forest (Info)	0.9978	0.9692	0.2243	0.9709	0.9842
Random Forest (Gini)	0.9956	0.97	0.3077	0.9739	0.9846
Support Vector Machine	0.9998	0.9645	0.0494	0.9647	0.9819
Multi-Layer Perceptron	0.9986	0.9663	0.1268	0.9674	0.9828

Table 20: Male Genital Cancer Over Sampling Results

Algorithms	Sensitivity	Accuracy	Specificity	Precision	F1-score
Decision Tree (C4.5)	0.9049	0.8923	0.5654	0.9818	0.9418
Decision Tree (Gini)	0.9	0.8876	0.5648	0.9817	0.9391
Logistic Regression	0.818	0.8154	0.7497	0.9884	0.8951
Naïve Bayes	0.8595	0.8445	0.4533	0.9761	0.9141
Random Forest (Info)	0.9158	0.9071	0.6802	0.9867	0.95
Random Forest (Gini)	0.9556	0.9414	0.5721	0.983	0.9691
Support Vector Machine	0.8357	0.8322	0.7403	0.9882	0.9056
Multi-Layer Perceptron	0.671	0.6735	0.7377	0.9852	0.7983

Table 21: Male Genital Cancer Under Sampling Results

Algorithms	Sensitivity	Accuracy	Specificity	Precision	F1-score
Decision Tree (C4.5)	0.7856	0.7847	0.7623	0.9885	0.8754
Decision Tree (Gini)	0.781	0.7809	0.7784	0.9892	0.8729
Logistic Regression	0.8136	0.8112	0.7477	0.9882	0.8924
Naïve Bayes	0.8821	0.865	0.4226	0.9754	0.9264
Random Forest (Info)	0.8451	0.8426	0.7757	0.9899	0.9118
Random Forest (Gini)	0.9479	0.9317	0.5721	0.98	0.9637
Support Vector Machine	0.8313	0.8282	0.7477	0.9884	0.9031
Multi-Layer Perceptron	0.6766	0.6754	0.6455	0.9802	0.8006

5.5 FEMALE GENITAL CANCER

The results of Female Genital Cancer indicated that oversampling and undersampling techniques have improved the specificity and precision of the models. Random Forests have proven to be the best predictors of all models, Support Vector Machines came as second-best predictors, then Logistic regression, Multi-Layer Perceptron, and Decision Trees. Naive Bayes turned out to be the least predictor of all models in case of no sampling technique (Table 22). In case of oversampling & undersampling techniques (Table 23 & Table 24), Random Forests have outperformed all other machine learning models. Support Vector Machines and Logistic Regression came

out to be second best predictors, then Decision Trees, Multi-Layer Perceptron, and Naive Bayes as usual with least prediction performance.

Table 22: Female Genital Cancer No Sampling Results

Algorithms	Sensitivity	Accuracy	Specificity	Precision	F1-score
Decision Tree (C4.5)	0.898	0.8432	0.6889	0.8904	0.8942
Decision Tree (Gini)	0.9004	0.8436	0.6841	0.8891	0.8947
Logistic Regression	0.9181	0.8573	0.686	0.8916	0.9047
Naïve Bayes	0.5972	0.6683	0.8683	0.9273	0.7265
Random Forest (Info)	0.9256	0.8664	0.6996	0.8966	0.9109
Random Forest (Gini)	0.9252	0.8668	0.7025	0.8975	0.9111
Support Vector Machine	0.9135	0.8576	0.7002	0.8955	0.9044
Multi-Layer Perceptron	0.893	0.8424	0.7002	0.8934	0.8932

Table 23: Female Genital Cancer Over Sampling Results

Algorithms	Sensitivity	Accuracy	Specificity	Precision	F1-score
Decision Tree (C4.5)	0.8297	0.8142	0.7706	0.9105	0.8683
Decision Tree (Gini)	0.8277	0.8153	0.7805	0.9139	0.8686
Logistic Regression	0.8299	0.8309	0.8336	0.9335	0.8787
Naïve Bayes	0.4607	0.5745	0.8947	0.9249	0.615
Random Forest (Info)	0.8501	0.8476	0.8408	0.9376	0.8917
Random Forest (Gini)	0.8575	0.8506	0.8314	0.9347	0.8944
Support Vector Machine	0.83	0.8321	0.838	0.9351	0.8794
Multi-Layer Perceptron	0.8616	0.8263	0.7268	0.8987	0.8798

Table 24: Female Genital Cancer Under Sampling Results

Algorithms	Sensitivity	Accuracy	Specificity	Precision	F1-score
Decision Tree (C4.5)	0.8018	0.8106	0.8356	0.9321	0.862
Decision Tree (Gini)	0.8036	0.811	0.8317	0.9307	0.8625
Logistic Regression	0.8309	0.8315	0.833	0.9333	0.8792
Naïve Bayes	0.4734	0.5834	0.8929	0.9256	0.6264
Random Forest (Info)	0.8376	0.8412	0.8513	0.9406	0.8861
Random Forest (Gini)	0.838	0.8421	0.8536	0.9415	0.8868
Support Vector Machine	0.83	0.832	0.8376	0.935	0.8794
Multi-Layer Perceptron	0.8117	0.798	0.7597	0.9048	0.8557

5.6 RESPIRATORY CANCER

In case of Respiratory Cancer, implementing sampling techniques have improved the sensitivity of the machine learning models, thereby improving their performance too. For no sampling technique (Table 25), Random Forests came as the best predictors of all other models, Logistic Regression as the second-best predictor, then Support Vector Machine, Decision Trees, Multi-Layer Perceptron and finally Naive Bayes with least performance. After implementing sampling techniques (Table 26 & Table 27), Random Forests turned out to be best predictors. Support Vector Machine and Logistic Regression were on par and came out to be second best predictors, then Decision Trees, Multi-Layer Perceptron, and Naive Bayes as usual turned out to be the least performing model of all.

Table 25: Respiratory Cancer No Sampling Results

Algorithms	Sensitivity	Accuracy	Specificity	Precision	F1-score
Decision Tree (C4.5)	0.6789	0.8802	0.9366	0.75	0.7127
Decision Tree (Gini)	0.6747	0.879	0.9362	0.7477	0.7093
Logistic Regression	0.6822	0.8876	0.9451	0.7768	0.7264
Naïve Bayes	0.4764	0.8512	0.9562	0.7529	0.5836
Random Forest (Info)	0.7165	0.895	0.9449	0.7847	0.7491
Random Forest (Gini)	0.7165	0.8951	0.9452	0.7854	0.7494
Support Vector Machine	0.6699	0.8876	0.9485	0.7848	0.7228
Multi-Layer Perceptron	0.61	0.8767	0.9514	0.7785	0.684

Table 26: Respiratory Cancer Over Sampling Results

Algorithms	Sensitivity	Accuracy	Specificity	Precision	F1-score
Decision Tree (C4.5)	0.7864	0.8513	0.8695	0.6279	0.6983
Decision Tree (Gini)	0.7834	0.8484	0.8665	0.6218	0.6933
Logistic Regression	0.8503	0.8596	0.8622	0.6336	0.7261
Naïve Bayes	0.4787	0.8512	0.9556	0.7511	0.5847
Random Forest (Info)	0.868	0.8674	0.8673	0.6469	0.7413
Random Forest (Gini)	0.8631	0.8705	0.8726	0.6549	0.7447
Support Vector Machine	0.8559	0.8599	0.861	0.633	0.7277
Multi-Layer Perceptron	0.8724	0.7961	0.7747	0.5203	0.6519

Table 27: Respiratory Cancer Under Sampling Results

Algorithms	Sensitivity	Accuracy	Specificity	Precision	F1-score
Decision Tree (C4.5)	0.8426	0.8512	0.8536	0.6172	0.7125
Decision Tree (Gini)	0.8448	0.8485	0.8495	0.6113	0.7093
Logistic Regression	0.851	0.8595	0.8619	0.6331	0.726
Naïve Bayes	0.4772	0.8512	0.9559	0.7521	0.5839
Random Forest (Info)	0.8728	0.8649	0.8626	0.6403	0.7387
Random Forest (Gini)	0.8734	0.8648	0.8624	0.64	0.7387
Support Vector Machine	0.8563	0.8597	0.8606	0.6324	0.7275
Multi-Layer Perceptron	0.695	0.8596	0.9058	0.6738	0.6843

5.7 URINARY CANCER

The results for Urinary Cancer showed that specificity of the models in case of no sampling technique are lower than acceptable ranges, so implementing oversampling and undersampling techniques improved the specificity of these models. For no sampling technique (Table 28), Random Forests came as the best predictor, Support Vector Machine came out to be the second-best predictor, then Logistic Regression, Decision Trees and Naive Bayes. Multi-Layer Perceptron has the least performance. After implementing sampling techniques (Table 29 & Table 30), results indicated that Random Forests turned out to be best predictors, Support Vector Machine as the second-best predictor then Logistic Regression, Decision Trees, and Multi-Layer Perceptron. Naive Bayes as usual turned out to be with least performance.

Table 28: Urinary Cancer No Sampling Results

Algorithms	Sensitivity	Accuracy	Specificity	Precision	F1-score
Decision Tree (C4.5)	0.925	0.8755	0.7307	0.9095	0.9172
Decision Tree (Gini)	0.9254	0.875	0.7275	0.9085	0.9169
Logistic Regression	0.9415	0.8819	0.7072	0.9039	0.9223
Naïve Bayes	0.8618	0.836	0.7606	0.9133	0.8868
Random Forest (Info)	0.9493	0.8919	0.7237	0.9095	0.929
Random Forest (Gini)	0.9502	0.8923	0.7228	0.9093	0.9293
Support Vector Machine	0.9407	0.881	0.7061	0.9035	0.9217
Multi-Layer Perceptron	0.8949	0.8047	0.5409	0.8508	0.8723

Table 29: Urinary Cancer Over Sampling Results

Algorithms	Sensitivity	Accuracy	Specificity	Precision	F1-score
Decision Tree (C4.5)	0.8587	0.8437	0.7996	0.9261	0.8912
Decision Tree (Gini)	0.8574	0.8445	0.807	0.9285	0.8915
Logistic Regression	0.8668	0.8585	0.8343	0.9387	0.9013
Naïve Bayes	0.8219	0.8151	0.7952	0.9215	0.8689
Random Forest (Info)	0.8812	0.8734	0.8506	0.9452	0.9121
Random Forest (Gini)	0.8885	0.8766	0.842	0.9427	0.9148
Support Vector Machine	0.8726	0.8631	0.8353	0.9394	0.9048
Multi-Layer Perceptron	0.7764	0.7941	0.8457	0.9364	0.8489

Table 30: Urinary Cancer Under Sampling Results

Algorithms	Sensitivity	Accuracy	Specificity	Precision	F1-score
Decision Tree (C4.5)	0.8398	0.8416	0.8467	0.9413	0.8877
Decision Tree (Gini)	0.8388	0.8418	0.8509	0.9427	0.8877
Logistic Regression	0.8672	0.8591	0.8357	0.939	0.9017
Naïve Bayes	0.8213	0.8147	0.7952	0.9215	0.8685
Random Forest (Info)	0.8694	0.8669	0.8596	0.9477	0.9068
Random Forest (Gini)	0.8716	0.8686	0.8597	0.9478	0.9081
Support Vector Machine	0.8732	0.8637	0.8359	0.9396	0.9052
Multi-Layer Perceptron	0.8387	0.8282	0.7976	0.9238	0.8792

Heatmaps below gives information on performance of the models across cancer cohorts.

Figure 8: F1 Score Heatmap for No Sampling Technique

Figure 9: F1 Score Heatmap for Over Sampling Technique

Figure 10: F1 Score Heatmap for Under Sampling Technique

Figure 11: AUC Heatmap for No Sampling Technique

Figure 12: AUC Heatmap for Over Sampling Technique

Figure 13: AUC Heatmap for Under Sampling Technique

5.8 DISCUSSION

For prediction of cancer survivability, one of the important components is quality of the data. All the data obtained cannot be used for survival prediction as it contains a lot of heterogeneous information like clinically irrelevant values, missing values, duplicate variables, deprecated variables etc. All these affect the survival prediction, so the data has to be cleaned and data relevant for survival prediction has to be chosen properly. After careful data preprocessing steps like variable elimination, data mapping, elimination of records with clinically irrelevant values and missing data, we chose a total of 21 independent variables which are relevant to survival prediction. For any prediction outcome, there has to be a target variable. In this study, we chose to predict 5-year survival since for cancer 5-years is chosen as standard survival time. A binary target variable is created based on certain conditions and choosing certain variables like survival months, type of follow-up, cause of death, vital status etc. After data preprocessing steps, machine learning models (Figure 6) were implemented along with 10-fold cross validation. From the results, we observed that the models were leaning towards either positive class or negative class. Later it was learned that models tend to bias towards majority class even after 10-fold cross validation was implemented, this is due to the datasets which are obtained after data preprocessing are highly imbalanced. Machine learning models tend to perform better in case of balanced classes otherwise they tend to bias towards the majority class. In order to remove this bias, we opted for class balancing techniques like oversampling and undersampling. The sampling techniques have shown to not only balance the classes, but also improve the performance of the models. In order to prove that class balancing techniques have improved the performance of the models, the experiment

was designed in a way that within a cancer cohort all the machine learning models will be implemented across all the sampling techniques and compared with each other and this process was carried out across all cancer cohorts. For comparing the performance of the models, we chose to compute metrics like sensitivity, accuracy, specificity, precision, f1 score and auc score. From the results obtained, it is evident that Random Forests have outperformed all other machine learning models in most cases and the reason for that might be, since random forests create many decision trees at training phase and selects mode of classes from all individual trees constructed. Support Vector Machines came out to be second, then Logistic Regression as third, Decision Trees in fourth place, Multi-Layer Perceptron in fifth place and finally Naive Bayes turned out to be the least performing model. The results have also shown that sampling techniques have improved the performance of the models. Overall, oversampling technique has been chosen for performance improvement because it has shown significant improvement that no sampling technique results and in case of undersampling, it cannot be chosen as it eliminates the data which might be useful for prediction purposes.

5.9 LIMITATIONS

Although the performance of the models is excellent and are in acceptable ranges, there are certain limitations to this project.

1. The data obtained from SEER do not have any information on cancer genes.
Gene information in prediction of cancer survival plays an important role.
2. SEER data do not provide information on dietary routine, drinking and smoking habits of the patient since they also play an important role in cancer survival.

3. Lymphoma and Leukemia Cancer cohort has been excluded from the study due to a smaller number of records.
4. Other Cancer cohort has been excluded from the study due to granularity of the data.
5. The current machine learning model can only predict the survival outcome of a case, but it cannot provide information on the factors that are affecting most on the individual case.
6. Machine learning models can only suggest for diagnosis and treatment alternatives, but it cannot replace the physician.
7. Human intervention is necessary to explore the knowledge extracted from the data since all the hidden patterns obtained are not useful and logical.

5.10 FUTURE WORK

The Future work in this project can be done in the following areas. Exploring the limitations of the project and develop new strategies. Exploring other methodologies and sampling techniques to improve the performance of the model and its behavior. Implementing the deep learning techniques for prediction of cancer survivability. Since the results are promising, developing this project into an application so that it can be used in point of care at a hospital organization.

BIBLIOGRAPHY

- 1 What Is Cancer? - National Cancer Institute. 2007.<https://www.cancer.gov/about-cancer/understanding/what-is-cancer>
- 2 Cancer of Any Site - Cancer Stat Facts. SEER. 2020.<https://seer.cancer.gov/statfacts/html/all.html>
- 3 Benard VB, Watson M, Saraiya M, *et al.* Cervical cancer survival in the United States by race and stage (2001-2009): Findings from the CONCORD-2 study. *Cancer* 2017;**123 Suppl**:5119–37. doi:10.1002/cncr.30906
- 4 Joseph DA, Johnson CJ, White A, *et al.* Rectal cancer survival in the United States by race and stage, 2001 to 2009: Findings from the CONCORD-2 study. *Cancer* 2017;**123 Suppl**:5037–58. doi:10.1002/cncr.30882
- 5 Steele CB, Li J, Huang B, *et al.* Prostate cancer survival in the United States by race and stage (2001-2009): Findings from the CONCORD-2 study. *Cancer* 2017;**123 Suppl**:5160–77. doi:10.1002/cncr.31026
- 6 Cancer Survival in the United States | CDC. 2019.<https://www.cdc.gov/cancer/dcpc/research/articles/concord-2-supplement.htm>
- 7 Rokach L, Maimon OZ. *Data Mining with Decision Trees: Theory and Applications*. World Scientific 2008. <https://www.google.com/books?id=8B9pDQAAQBAJ>
- 8 Lavrac N. Selected techniques for data mining in medicine. *Artif Intell Med* 1999;**16**:3–23. doi:10.1016/s0933-3657(98)00062-1
- 9 CDC Media Relations: Press Release: Progress Shown in Death Rates from Four Leading Cancers. 2020.<https://www.cdc.gov/media/pressrel/r030902.htm>
- 10 Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. *Artif Intell Med* 2005;**34**:113–27. doi:10.1016/j.artmed.2004.07.002
- 11 García-Laencina PJ, Abreu PH, Abreu MH, *et al.* Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values. *Comput Biol Med* 2015;**59**:125–33. doi:10.1016/j.compbimed.2015.02.006
- 12 Shukla N, Hagenbuchner M, Win KT, *et al.* Breast cancer data analysis for survivability studies and prediction. *Comput Methods Programs Biomed* 2018;**155**:199–208. doi:10.1016/j.cmpb.2017.12.011
- 13 Al-Bahrani R, Agrawal A, Choudhary A. Colon cancer survival prediction using ensemble data mining on SEER data. In: *2013 IEEE International Conference on Big Data*. IEEE 2013. 9–16. doi:10.1109/BigData.2013.6691752

- 14 Agrawal A, Misra S, Narayanan R, *et al.* Lung Cancer Survival Prediction using Ensemble Data Mining on Seer Data. *Sci Program* 2012;**20**:29–42. doi:10.1155/2012/920245
- 15 Hegselmann S, Greulich L, Varghese J, *et al.* Reproducible Survival Prediction with SEER Cancer Data. *Mlhc-2018* 2018;;26.files/950/Hegselmann et al. - Reproducible Survival Prediction with SEER Cancer .pdf
- 16 Sarvestani AS, Safavi AA, Parandeh NM, *et al.* Predicting breast cancer survivability using data mining techniques. In: *2010 2nd International Conference on Software Technology and Engineering (ICSTE 2010)*. IEEE 2010. 5608818. doi:10.1109/ICSTE.2010.5608818
- 17 Kate RJ, Nadig R. Stage-specific predictive models for breast cancer survivability. *Int J Med Inform* 2017;**97**:304–11. doi:10.1016/j.ijmedinf.2016.11.001
- 18 Delen D. Analysis of cancer data: a data mining approach. *Expert Syst* 2009;**26**:100–12. doi:10.1111/j.1468-0394.2008.00480.x
- 19 Dimitoglou G, Adams J, Jim C. Comparison of the C4.5 and a Naive Bayes Classifier for the Prediction of Lung Cancer Survivability. 2012;**4**.<http://files/1095/Dimitoglou et al. - 2012 - Comparison of the C4.5 and a Naive Bayes Classifie.pdf>
- 20 Silva A, Oliveira T, Neves J, *et al.* Treating Colon Cancer Survivability Prediction as a Classification Problem. *ADCAIJ Adv Distrib Comput Artif Intell J* 2016;**5**:37. doi:10.14201/ADCAIJ2016513750
- 21 Submit Request For Data - SEER Datasets. SEER. 2020.<https://seer.cancer.gov/seertrack/data/request/>
- 22 SEER RESEARCH DATA RECORD DESCRIPTION 2018. 2018;;86.files/1108/2018 - SEER RESEARCH DATA RECORD DESCRIPTION 2018.pdf
- 23 Hu Z, Melton GB, Arsoniadis EG, *et al.* Strategies for handling missing clinical data for automated surgical site infection detection from the electronic health record. *J Biomed Inform* 2017;**68**:112–20. doi:10.1016/j.jbi.2017.03.009
- 24 Santos MS, Soares JP, Abreu PH, *et al.* Cross-Validation for Imbalanced Datasets: Avoiding Overoptimistic and Overfitting Approaches [Research Frontier]. *IEEE Comput Intell Mag* 2018;**13**:59–76. doi:10.1109/MCI.2018.2866730
- 25 Xu Z, Feng Y, Li Y, *et al.* Predictive Modeling of the Risk of Acute Kidney Injury in Critical Care: A Systematic Investigation of The Class Imbalance Problem. *AMIA Summits Transl Sci Proc* 2019;**2019**:809–18.<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6568062/>
- 26 Fotouhi S, Asadi S, Kattan MW. A comprehensive data level analysis for cancer diagnosis on imbalanced data. *J Biomed Inform* 2019;**90**:103089. doi:10.1016/j.jbi.2018.12.003