5-2021

Machine Learning Approach To Predict Mortality Rates Based On Hospital Clinical Data

Rebecca Smith

Machine Learning Approach to Predict Mortality Rates Based on Hospital Clinical Data

by

Rebecca Smith

Bachelor of Science
Mathematics
Florida Southern College
2019

A thesis
submitted to the College of Engineering and Science
at Florida Institute of Technology
in partial fulfillment of the requirements
for the degree of

Master of Science
in
Operations Research

Melbourne, Florida
May, 2021

We the undersigned committee
hereby approve the attached thesis Machine Learning Approach to

Predict Mortality Rates Based on Hospital Clinical Data
by Rebecca Smith

Munevver Mine Subasi, Ph.D.
Associate Professor and Department Head
Mathematical Sciences
Committee Chair

Joo Young Park, Ph.D.
Assistant Professor
Education and Interdisciplinary Studies
Outside Committee Member

Ryan White, Ph.D.
Assistant Professor
Mathematical Sciences
Committee Member

Susan Serrano, Ph.D.
Associate Professor
Mathematics, Florida Southern College
Guest Member

ABSTRACT

Title:

Machine Learning Approach to Predict Mortality Rates Based on Hospital Clinical Data

Author:

Rebecca Smith

Major Advisor:

Munevver Mine Subasi, Ph.D.

This thesis integrates fundamental concepts from conventional statistics with the more explanatory, algorithmic, and computational techniques offered by machine learning to predict early mortality risk of surgical patients. Well-known classification methods, including Random Forest, Decision Trees, Nearest Neighbor, Stochastic Gradient Descent, Logistic Regression, Naïve Bayes, Bayes Network, Neural Networks, and Support Vector Machines, are utilized to predict mortality risk of elective general surgical patients treated between January 2005 and September 2010 at the Cleveland Clinic [33]. Clinical factors include surgery type, age, gender, race, BMI, underlying chronic conditions, surgical risk indices, surgical timing predictors, the 30-day mortality, and in-hospital complication for each patient. 10×10-folding cross validation experiments are conducted to evaluate the prediction performance on low, medium, and high mortality risk groups. A Decision Tree classification model consisting of 83 low and 135 high risk patterns is presented. The overall average accuracy of the classifiers applied to predict low and high risk mortality is 85.2% with precision of 0.89, recall of 0.95, and F-measure of 0.92. The overall accuracy of the classifiers applied to predict low, medium, and high risk mortality is 84.7% with precision of 0.89, recall of 0.94, and F-measure of 0.91.

Table of Contents

List of Figures

List of Tables

List of Symbols, Nomenclature or Abbreviations

kNN	k-Nearest Neighbors
RF	Random Forests
SVM	Support Vector Machines
SVC	Support Vector Classifiers
SVR	Support Vector Regression
SMO	Sequential Minimal Optimization
MLP	Multilayer Perceptron
CVD	Cardiovascular Disease
Osteoart	Osteoarthritis
Psych	Psychiatric Disorder
RSI	Risk Stratification Index
ASA	American Society of Anesthesiologist
BMI	Body Mass Index

Acknowledgements

I would like to start off by thanking my research and academic advisor Dr. Munevver Mine Subasi, Associate Professor and Department Head of Mathematical Sciences. Without her help and support throughout this process, I would not have been able to complete the work on this research. I am sure that when she met me at the 2019 ASPIRE Conference hosted by Florida Gulf Coast University and gave me advice on my project that she did not expect I would be coming to her college to complete my Master's degree under her.

I would also like to thank each of my committee members, Dr. Ryan White, Dr. Susan Serrano, Dr. Joo Young Park, and Dr. Munevver Mine Subasi for being a sounding board for my ideas and for giving me such helpful suggestions. I also want to thank Florida Institute of Technology for the opportunity to continue my education and doing the research that I have come to love. I also would like to commend the school for the facilities that are available to the students and for the academic quality of each of the courses. I especially want to state that the professors at this university show a genuine interest in the students and want to help them succeed in any way possible.

Dedication

I would like to dedicate this thesis to my family who has supported me through this entire journey. They have been there through the happy and tough times and have given me the support that I needed to keep on going and reach my goals. I would also like to dedicate this thesis to Dr. Susan Serrano who introduced me to statistics and my love of research. I appreciate all the time that you took to teach me and guide me through my first research project as an undergraduate student. Finally, I would like to dedicate this to Dr. Roxanne Back who helped to push me to keep going even when I thought that I could not reach my goals. I would just like to thank everyone who has pushed me to be the person that I am today. I would not be here without all the love and support they have given me.

Chapter 1

Introduction

Early mortality, defined as 30-day mortality, after surgery received considerable attention in literature [6, 20, 21, 22, 33, 34]. These studies aim at predicting the mortality risk or identifying survivors and nonsurvivors based on the clinical features of surgery patient. It was shown that human factors such as fatigue, surgery schedule, and surgery staff had an impact on mortality risk of surgery patients. Sessler et al. (2011)[33] investigated impact of surgery schedule on the risk of 30-day mortality associated with elective general surgery. As part of the study, Sessler et al. (2011) [33] collected "Surgery Timing" dataset containing 32,001 elective general surgical patients treated between January 2005 and September 2010 at the Cleveland Clinic. In addition to the surgical timing predictors such as hour, day of week, month, moon phase, the clinical feature included surgery type, age, gender, race, BMI, underlying chronic conditions, surgical risk indices as well as the 30-day mortality and in-hospital complication for each patient.

The primary outcome of "Surgery Timing" study conducted by Sessler et al. (2011) [33] was all-cause 30-day mortality obtained from a review of hospital

records and the Social Security Death Index database. Sessler et al. (2011) [33] modeled the 30-day mortality using multivariable logistic regression. In the same study, Sessler et al. (2011) [33] considered the composite complications defined by United States Agency for Healthcare Research and Quality's Clinical Classifications Software (AHRQ-CCS) diagnosis categories 237 of which were complication of device, implant or graft and 238 of which were complications of surgical procedures or medical care.

Sessler et al. (2011) [33] adjusted for diagnoses and procedures using the Risk Stratification Index (RSI) for 30-day mortality. *Mortality RSI* of the Cleveland Clinic surgery patients was obtained to predict 3-day mortality from the International Classification of Diseases, 9th revision, Clinical Modification (ICD-9-CM). 30-day mortality and in-hospital complications of the Clevend Clinic surgery patients were modeled using multivariable logistic regression that provided the adjucted incidence of 30-day mortality and in-hospital complication based on hour, day, month of the surgery and moon phase when the surgical procedure was started.

In this study, we adopt the more explanatory, algorithmic, and computational techniques offered by machine learning to stratify surgery patients into low, medium, and high mortality risk groups and identify the clinical features for mortality risk of patients in "Surgery Timing" dataset.

The organization of this thesis is as follows. Chapter 2 briefly outlines well-known and commonly used machine learning methods and metrics that can be used to evaluate the prediction power of classification methods. describes the study subjects and preprocessing of Surgery Timing dataset. Chapter 3 describes the study subjects and provides the statistics for clinical features used in Surgery Timing study. Identification of low risk, medium risk, and high risk are presented

in Chapter 4. Overall performance of well-known classification methods and combinatorial patterns that can be used to predict the mortality risk groups are also presented in Chapter 4. Finally, in Chapter 5, the discussion concludes with a summary of data analysis results.

Chapter 2

Review of Classification Methods

2.1 Machine Learning

Machine Learning is a data analysis method that focuses on building applications learned from data and automates analytical model building to improve predictions through experience. The fundamental concepts from conventional statistics and optimization are integrated with machine learning techniques to develop systematic procedures to analyze large-scale, complex structured datasets generated by sophisticated technologies used in science and engineering. A typical data analysis process comprises four phases [2, 4]:

- data preprocessing, including data transformation, imputation, feature selection, and dimensionality reduction,
- class discovery (unsupervised learning) or classs comparison and discrimination (supervised learning),
- evaluation (statistical tests or cross-validation) of the prediction, and

- interpretation of the results.

Unsupervised learning uses clustering analysis to identify subgroups of observations in datasets with no known outcome. Supervised learning is a machine learning technique that learns from the data predict continuous valued outcome (regression analysis), discrete valued outcome (classification), and time-to-event outcome (survival analysis).

Our study focuses on stratifying surgery patients into different groups based on their mortality risk. To achieve this, we adopt well-known and commonly used supervised learning classification methods, including Logistic Regression, Decision Trees, Random Forest, Naïve Bayes, Stochastic Gradient Descent, Nearest Neighbor, Neural Networks, and Support Vector Machines. Below we briefly outline these methods.

2.1.1 Logistic Regression

Logistic Regression is a parametric classification model that is used to predict the discrete outcome in multivariate data. The method uses the weighted sum given in Equation (2.1):

$$X = \beta_0 + \sum_{i=1}^{n} \beta_k x_k \tag{2.1}$$

where $\beta_k, k = 0, 1, ..., n$ are parameters. The weighted sum is used in Sigmoid function given in Equation (2.2) to calculate the probability of the input being in a specific category.

$$Sigmoid(X) = \frac{1}{1 + e^{-(\beta_0 + \sum_{i=1}^{n} \beta_k X_k)}} \tag{2.2}$$

Logistic Regression uses the log odds ratio and an iterative maximum likelihood

to fit the final model. It is relatively efficient classification method.

2.1.2 Decision Tree and Random Forest

Decision Tree, illustrated in Figure 2.1, is a classification method that uses tree-like models containing explicit decision rules that can predict discrete valued outcomes [5, 9]. J48 (C4.5) Decision Tree algorithm [29] is an extension of Iterative Dichotomiser 3 algorithm. Although over-fitting is common, Decision Tree is often adopted in data analysis due to its high interpretability and intuitive nature.

Random Forest method builds a large collection of de-correlated decision trees and report the average predictions of the decision trees generated [7]. The method can also be referred to as average tree estimator. Random Forest uses bagging technique to minimize over-fitting.

2.1.3 Bayesian Network and Naïve Bayes Method

Bayesian Networks are built off the idea of Bayes' theorem [12]. The Bayesian Network is a directed acyclic graphs that allows efficient and effective representation of the joint probability distribution variables in data. Nodes in the network represent random variables and edges represent directed correlations between the variables, where nodes are assumed to be conditionally independent of the parent nodes. It uses the conditional probability

$$P[Cause|Evidence] = P[Evidence|Cause] * \frac{P[Cause]}{P[Evidence]} \tag{2.3}$$

Naïve Bayes is another probabilistic classifier [5, 9, 24] that also uses Bayes' theorem. The method is called naive because it assumes that the features are

independent random variables. Let $P(X|c)$ be the probability of the predictor for a paricular class c. Then given that the predictor of X, the probability of assigning class c, $P(c|X)$, is defined by

$$P(c|X) = P(X_1|c)P(X_2|c)...P(X_n|c)P(c) \tag{2.4}$$

where $P(c)$ is the prior probability of the class and $P(X)$ is the prior probability of the predictor.

Naïve Bayes is a computationally inexpensive method which performs well if the input dataset indeed contains independent features.

2.1.4 Stochastic Gradient Descent

Stochastic Gradient Descent classifier [1] is an iterative algorithm that implements stochastic gradient descent method for learning various linear models. The method consists of six steps [36]:

- find the slope of the objective function also known as finding the gradient of the function,
- pick a random initial value for each of the parameters,
- update the gradient function by plugging in the parameter values,
- calculate the step sizes for each feature using the following equation:

$$stepsize = gradient * learningrate, \tag{2.5}$$

- calculate the new parameters

$$newparams = oldparams - stepsize \tag{2.6}$$

- repeat steps three to five until the gradient is close to 0.

2.1.5 k-Nearest Neighbors

k-Nearest Neighbors, commonly known as kNN, is an algorithm that classifies observations based on the distance between them [23]. The algorithm uses a hyperparameter k that represents the number of neighbors. Class of an observation is determined based on the most common classes (closest distance) among the observation's neighbors. Nearest Neighbor is an instance-based learning method and assumes that the distance between the observations is sufficient enough to make an inference about the observation to be predicted.

2.1.6 Multilayer Perceptron

Multilayer Perceptron is a feed forward neural network with multiple layers [10]. For example, in a 3-layer network, the first layer would be the input layer, the second would be the hidden layer and the final layer would be the output layer. The number of hidden layers is determined by the user. Feed forward neural network assumes all of nodes are fully connected (i.e., it is a complete graph) as illustrated in Figure 2.2.

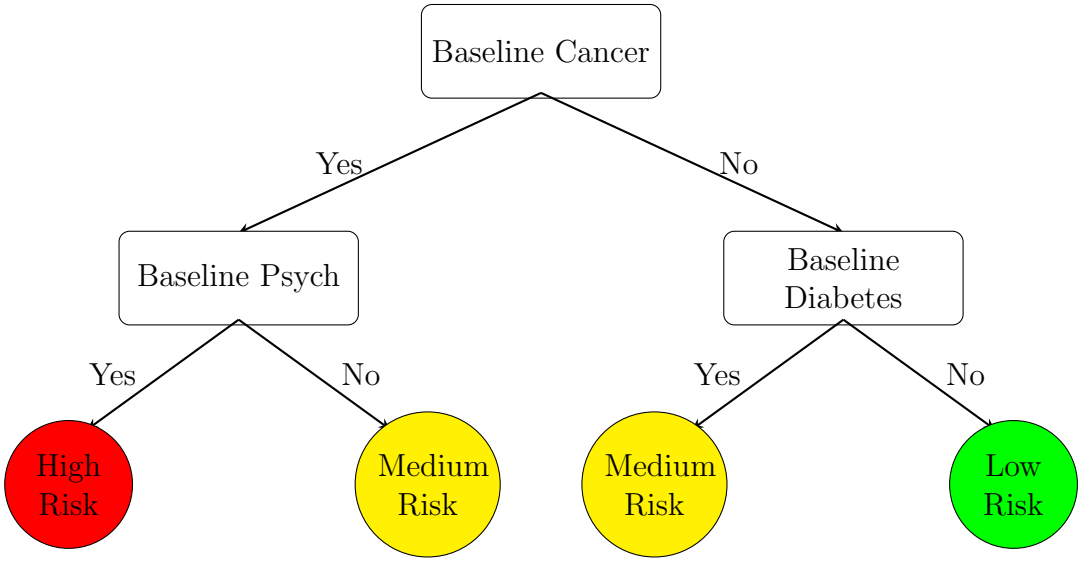

Figure 2.1: Illustration of Decision Tree

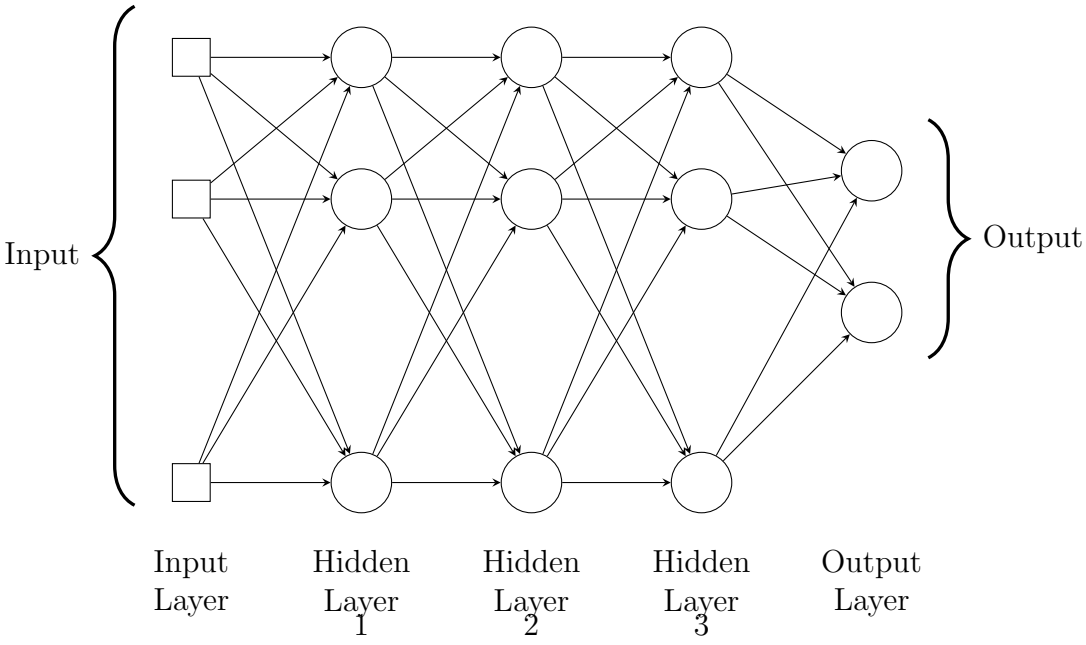

Figure 2.2: Illustration of Multi-layer Perceptron Neural Network

2.1.7 Support Vector Machines

Support Vector Machines (SVM) [8, 31] can be used for both regression and classification problems: Support Vector Regression and Support Vector Classification. The method find a hyperplane that can separate the data into different classes where the margin of the separation is maximized. Observations in different subgroups closest to the hyperplane are called the support vectors. The method then aims at maximizing the distance between the support vectors and the hyperplane.

Sequential Minimal Optimization (SMO) is an implementation of SVM algorithm that does not require the use of any more matrix storage and numerical quadratic programming steps. The method chooses two Lagrange multipliers and analytically optimizes the multipliers, avoiding quadratic optimization.

2.2 Performance Evaluation Metrics for Classification Methods

2.2.1 Area Under ROC

The area under the ROC is referring to what all falls below the ROC curve. The ROC curve is an illustration that checks how the classifier is performing by looking at the true positive rate as the false positive rate is changing. Now when you look at this you want it to stand out and be higher in the top-left corner of the plot. Now where that higher point is located is where you can look to see what the area under ROC value is going to be.

2.2.2 Precision and Recall

Precision of a classification method is the proportion of correctly classified positive observations:

$$Precision = \frac{Number\ of\ True\ Positives}{Number\ of\ True\ Positives + Number\ of\ False\ Positives}. \tag{2.7}$$

Recall of a classification method is the proportion of positive observations that were correctly predicted:

$$Recall = \frac{Number\ of\ True\ Positives}{Number\ of\ True\ Positives + Number\ of\ False\ Negatives}. \tag{2.8}$$

For example, a precision of 0.8 means when the classifier assigns an observation to the positive class, it is correct 80% of the time. Similarly, a precision of 0.15 means, the classifier correctly identifies 15% of the positive observations. The higher the precision and recall are the better the prediction perfomance of the classifier is.

2.2.3 F-Measure

The F-measure is a weighted mean of precision (P) and recall (R) defined as

$$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} \tag{2.9}$$

where $\alpha \in [0, 1]$.

F-measure can be considered as compromise between precision and recall. When both precision and recall are high, the corresponding F-measure is closer to 1 which is considered as significant prediction.

2.2.4 Mean Absolute Error

Mean Absolute Error measure how far the predicted observations in test set are away from the observations in a specific class in training set. It is the average over the test sample of the absolute differences between predicted value and observed value. The smaller the value of mean absolute error is, the better the prediction of classes in the input dataset.

Chapter 3

Study Subjects & Data Preprocessing

3.1 Input Data

The goal of this study is to stratify elective general surgical patients into different risk groups based on clinical features. To achieve this goal we use the "Surgery Timing" dataset containing 32,001 elective general surgical patients treated between January 2005 and September 2010 at the Cleveland Clinic [33]. The clinical features include surgery type, age, gender, race, BMI, underlying chronic conditions, surgical risk indices, the surgical timing predictors such as hour, day of week, month, moon phase as well as the 30-day mortality and in-hospital complication for each patient as shown in Table 3.1.

Table 3.1: Clinical Features in Surgery Timing Dataset

$V1$	Surgery Type
$V2$	Age
$V3$	Gender
$V4$	Race
$V5$	ASA Status
$V6$	BMI
$V7$	Baseline Cancer
$V8$	Baseline CVD
$V9$	Baseline Dementia
$V10$	Baseline Diabetes
$V11$	Baseline Digestive
$V12$	Baseline Osteoart
$V13$	Baseline Psych
$V14$	Baseline Pulmonary
$V15$	Baseline Charlson
$V16$	ccsMort30rate
$V17$	ccsComplicationRate
$V18$	Hour
$V19$	Day of Week
$V20$	Month
$V21$	Moon Phase
$V22$	mort30
$V23$	complication

Table 3.2: Surgical Procedures

Label	**Surgery Type**
A	Other
B	Arthroplasty Knee
C	Colorectal Resection
D	Endoscopy and Endoscopic Biopsy of the Urinary Tract
E	Gastrectomy; Partial and Total
F	Genitourinary Incontinence Procedures
G	Hip Replacement; Total and Partial
H	Hysterectomy; Abdominal and Vaginal
I	Inguinal and Femoral Hernia Repair
J	Laminectomy; Excision Intervertebral Disk
K	Lumpectomy; Quadrantectomy of Breast
L	Mastectomy
M	Nephrectomy; Partial and Complete
N	Oophorectomy; Unilateral and Bilateral
O	Open Prostatectomy
P	Other Excision of Cervix and Uterus
Q	Other Hernia Repair
R	Plastic Procedures on Nose
S	Repair of Cystocele and Rectocele; Obliteration of Vaginal Vault
T	Small Bowel Resection
U	Spinal Fusion
V	Thyroidectomy; Partial or Complete
W	Transurethral Resection of Prostates(TURP)

Below, we briefly outline the characteristics of clinical features in Surgery Timing dataset. The specific surgical procedures performed on Cleveland Clinic patients during the period of January 2005 to September 2010 are presented in Table 3.2. The characteristics of clinical features included in Surgery Timing dataset are shown in Table 3.3, where N is the number of patients and N^* is the number of patients with missing values in corresponding features.

$V1$**-Surgery Type:** Nominal valued feature representing the specific surgery that was performed.

$V2$**-Age:** Continuous valued feature representing patient's age at the time of surgery.

$V3$**-Gender:** Binary feature representing gender of patient: male (0), female (1).

$V4$**-Race:** Discrete values feature representing race of patient: Caucasian (1), African-American (2), other (3).

$V5$**-ASA Physical Status:** Categorical feature representing anesthesiologist physical status, where a value of 1 is assigned if the anesthesiologist of a surgery patient had a level of I-II, a value of 2 for level III, and a value of 3 for level IV-VI physical status.

$V6$**-BMI:** Continuous valued feature representing patient's body mass index (BMI) at the time of surgery.

$V7$**-Baseline Cancer:** Binary feature representing if patient has cancer (1) or not (0).

$V8$**-Baseline CVD:** Binary feature representing if patient has cardiovascular/cerebrovascular disease (1) or not (0).

$V9$**-Baseline Dementia:** Binary feature representing if patient has dementia (1) or not (0).

$V10$**-Baseline Diabetes:** Binary feature representing if patient is diabetic (1) or not (0).

$V11$**-Baseline Digestive:** Binary feature representing if patient has digestive disorder (1) or not (0).

$V12$**-Baseline Osteoart:** Binary feature representing if patient has osteoarthritis (1) or not (0).

$V13$**-Baseline Psych:** Binary feature representing if patient has psychiatric disorder (1) or not (0).

$V14$**-Baseline Pulmonary:** Binary feature representing if patient has pulmonary problems (1) or not (0).

$V15$**-Baseline Charlson:** Continuous valued feature representing the Charlson Comorbidity Index for each patient.

$V16$**-ccsMort30rate:** Nominal valued feature representing the overall incidence of 30-day mortality for each procedure category.

$V17$**-ccsComplicationRate:** Nominal valued feature representing the overall incidence of in-hospital complications for each procedure category.

$V18$**-Hour:** Discrete valued feature representing the specific hour that the procedure was performed. The values for this run from 1 to 24 with this being military time for each of the different hours throughout the day.

$V19$**-Day of Week:** Discrete valued feature representing the specific day on which the procedure was performed: Monday (1), Tuesday (2), Wednesday (3), Thursday (4), and Friday (5).

$V20$**-Month:** Discrete valued feature (1,...,12) representing the specific month that the procedure was performed.

$V21$**-Moonphase:** Discrete valued feature representing the moon phase in which the procedure has started: new moon (1), first quarter (2), full moon (3), and last quarter (4).

$V22$**-mort30:** Binary feature representing whether a patient experienced mortality within the first thirty days after the procedure (1) or not (0).

$V23$**-Complication:** Binary feature representing whether a patient experienced any complications while in the hospital (1) or not (0).

Table 3.3: Surgery Timing Dataset Characteristics

Feature	N	N^*	**Mean**	**StDev**	**Quartile 1**	**Median**	**Quartile 3**
Age	31999	2	57.66	15.04	48.2	58.6	68.3
Gender	31998	3	0.46	0.50	0	0	1
Race	31521	480	1.20	0.49	1	1	1
ASA Phsical Status	31993	8	1.49	0.56	1	1	2
BMI	28711	3290	29.45	7.27	24.6	28.19	32.81
Baseline_cancer	32001	0	0.34	0.47	0	0	1
Baseline_cvd	32001	0	0.51	0.50	0	1	1
Baseline_dementia	32001	0	0.01	0.09	0	0	0
Baseline_diabetes	32001	0	0.13	0.34	0	0	0
Baseline_digestive	32001	0	0.22	0.41	0	0	0
Baseline_osteoart	32001	0	0.18	0.38	0	0	0
Baseline_psych	32001	0	0.09	0.29	0	0	0
Baseline_pulmonary	32001	0	0.11	0.31	0	0	0
Baseline_Charlson	32001	0	1.18	1.88	0	0	2
Mortality_rsi	32001	0	-0.53	1.04	-1.24	-0.3	0
Complication_rsi	32001	0	-0.41	1.20	-0.84	-0.27	0
ccsMort30Rate	32001	0	0.00	0.00	0.000789	0.002764	0.007398
ccsComplicationRate	32001	0	0.13	0.09	0.08198	0.10937	0.18337
Hour	32001	0	10.38	2.92	7.65	9.65	12.72
Day of Week	32001	0	2.90	1.42	2	3	4
Month	32001	0	6.42	3.33	4	6	9
Moonphase	32001	0	2.52	1.11	2	3	4
mort30	32001	0	0.00	0.07	0	0	0
Complication	32001	0	0.13	0.34	0	0	0

The primary outcome of Surgery Timing study conducted by Sessler et al. (2011) [33] was all-cause 30-day mortality, $V22$-mort30, obtained from a review of hospital records and the Social Security Death Index database. Sessler et al. (2011) [33] modeled the 30-day mortality using multivariable logistic regression. In the same study, Sessler et al. (2011) [33] considered the composite complications defined by United States Agency for Healthcare Research and Quality's Clinical Classifications Software (AHRQ-CCS) diagnosis categories 237 of which were complication of device, implant or graft and 238 of which were complications of surgical procedures or medical care.

Sessler et al. (2011) [33] adjusted for diagnoses and procedures using the Risk Stratification Index (RSI) for 30-day mortality. *Mortality RSI*, shown in Figure 3.1, was obtained to predict 3-day mortality from the International Classification of Diseases, 9th revision, Clinical Modification (ICD-9-CM). Similarly, *Complication RSI*, shown in Figure 3.2, was made available in Surgery Timing dataset.

Mortality RSI values range from -4.4 to 4.86. The values are assigned based on the ICD-9-CM system to diagnoses and procedures associated with hospitals in the United States. *Mortality RSI* is equal to the logit of 30-day mortality defined as

$$logit = log(odds) = log(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1 + ...\beta_k X_k. \tag{3.1}$$

Then the probability that the patient dies within 30 days after the surgery is given by

$$p = \frac{e^{\beta_0+\beta_1 X_1+...\beta_k X_k}}{1+e^{\beta_0+\beta_1 X_1+...\beta_k X_k}}. \tag{3.2}$$

For example, a patient who has a mortality RSI value of 4.86 has the logit value of 4.86 and from Equation (3.2) the probability that the patient dies within

30 days after the surgical procedure is 0.9923:

$$p = \frac{e^{4.86}}{1 + e^{4.86}} = 0.9923 \tag{3.3}$$

Both *Mortality RSI* and *Complication RSI* are continuous valued outcomes. In this study, we aim at predicting *Mortality RSI* values of the patients in Surgery Timing dataset.

As an initial step, we find the correlation between *Mortality RSI* and each feature, including *Complication RSI*. As can be seen from Table 3.4, none of the features is significantly correlated with *Mortality RSI* whereas there is a significant positive correlation between *Mortality RSI* and *Complication RSI*. It is indeed expected to have a patient's mortality risk increase as the patient's in-hospital complication risk increases. In order to avoid any bias, we remove *Complication RSI* from the dataset to identify a classification model that can predict the mortality risk of patients in Surgery Timing dataset.

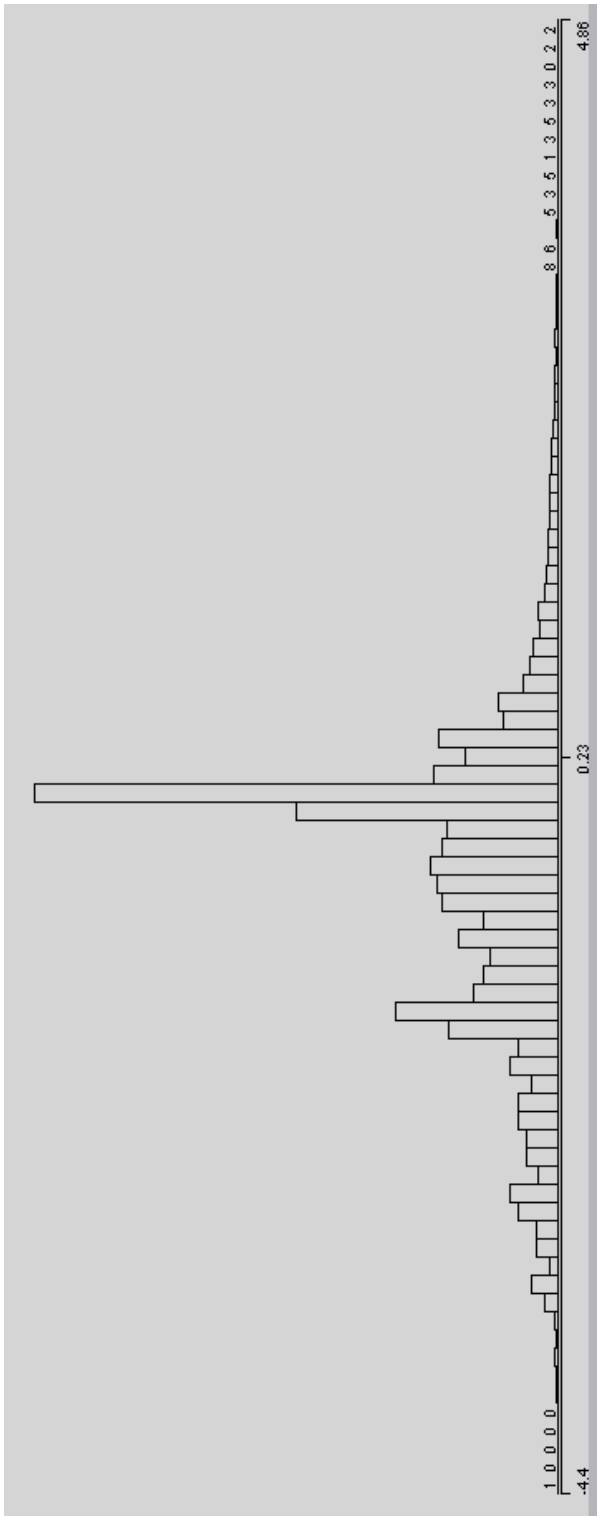

Figure 3.1: Mortality RSI for Surgery Timing Data

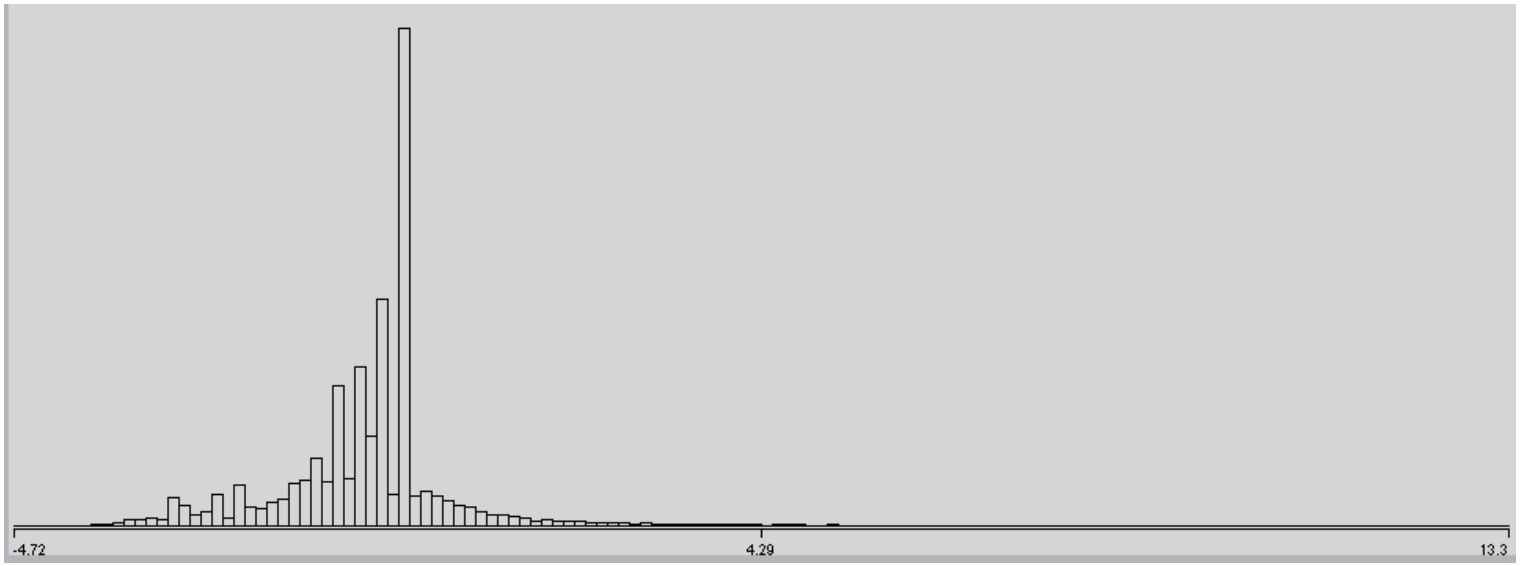

Figure 3.2: Complication RSI for Surgery Timing Data

Table 3.4: Correlations Between Mortality RSI and Other Features

Compared Features	**Correlation Value**
Mortality-RSI vs. Age	-0.099
Mortality-RSI vs. Gender	-0.019
Mortality-RSI vs. Race	0.033
Mortality-RSI vs. ASA-Status	0.064
Mortality-RSI vs. BMI	-0.108
Mortality-RSI vs. Baseline Cancer	0.198
Mortality-RSI vs. Baseline CVD	-0.132
Mortality-RSI vs. Baseline Dementia	0.054
Mortality-RSI vs. Baseline Diabetes	-0.052
Mortality-RSI vs. Baseline Digestive	-0.025
Mortality-RSI vs. Baseline Osteoart	-0.554
Mortality-RSI vs. Baseline Psych	-0.023
Mortality-RSI vs. Baseline Pulmonary	0.010
Mortality-RSI vs. Baseline Charlson	0.254
Mortality-RSI vs. Hour	0.122
Mortality-RSI vs. Day of Week	0.075
Mortality-RSI vs. Month	-0.002
Mortality-RSI vs. Moon Phase	-0.002
Mortality-RSI vs. Complication-RSI	0.723

3.2 Prediction of Mortality RSI

Our first attempt is to use regression analysis methods to predict *Mortality RSI* of 32,001 patients in Surgery Timing dataset. We run our analyses using WEKA data mining software [13]. Table 3.5 show the average of 10×10 folding cross-validation experiments for Linear Regression and Locally Weighted Naïve Bayes, respectively.

Based on 10×10-folding cross-validation experiments presented in Tables 3.5 and 3.6, we conclude that both Linear Regression and Locally Weighted Naïve Bayes have poor accuracy of predicting *Mortality RSI* (continuous valued outcome)

Table 3.5: Linear Regression Cross-Validation Results for *Mortality RSI*

Correlation coefficient	0.73
Mean absolute error	0.52
Root mean squared error	0.71
Relative absolute error	64.84%
Root relative squared error	68.45%

Table 3.6: Locally Weighted Naïve Bayes Cross-Validation Results for *Mortality RSI*

Correlation coefficient	0.61
Mean absolute error	0.62
Root mean squared error	0.82
Relative absolute error	77.14%
Root relative squared error	79.29%

of the patients in Surgery Timing dataset.

In this study, we adopt the more explanatory, algorithmic, and computational techniques offered by machine learning to stratify surgery patients into subgroups based on their low, medium, and high risk of mortality.

Chapter 4

Prediction of Mortality Risk in Surgery Patients

4.1 Prediction of Low and High Risk Mortality

4.1.1 Identification of Low and High Risk Patients

We recall that the values of *Mortality RSI* range from -4.4 to 4.86, i.e., the probability of patients dying within 30 days of surgery varies between 1.21% to 99.23% as calculated by Equation (3.2).

In this section, we aim at identifying two disjoint subgroups of Surgery Timing data, representing the high risk and low risk patients based on their *Mortality RSI* values. To achieve this we consider the two extreme ends of the *Mortality RSI* distribution presented in Figure 3.1.

- A patient is labeled as "low risk" patient if the patient's corresponding *Mortality RSI* value is between -4.4 and -1, i.e., the probability that the patient's

dies within 30 days after the surgery is between 1.21% and 26%.

- A patient is labeled as "high risk" patient if the patient's corresponding *Mortality RSI* value is between 1 and 4.86, i.e., the probability that the patient's dies within 30 days after the surgery is between 73.1% and 99.2%.

Patients whose *Mortality RSI* values fall into interval $(-1, 1)$ are removed from the dataset. The resulting dataset, referred to as "Surgery Timing LH", contains 9,559 low risk and 1,469 high risk patients and their corresponding clinical features in Table 3.1.

Table 4.1 and Figure 4.1 show the distribution of the low risk and high risk patients based on their surgical procedure. Table 4.2 and Figures 4.2-4.6 give the distribution of the low risk and high risk patients based on their age, gender, race, ASA physical status, and BMI, respectively. Distribution of underlying health conditions among low risk and high risk patients are presented in Table 4.3 and Figures 4.7-4.13. In Table 4.4 and Figures 4.14-4.16, we give the distribution of baseline Charlson index, ccsMort30rate, and ccsComplication rate among low risk and high risk patients, respectively. Table 4.5 and Figures 4.17-4.22 show the distribution of hour, day of week, month, moon phase, 30-day mortality and in-hospital complication of low risk and high risk patients, respectively.

Table 4.1: Low Risk and High Risk Patients Data Characteristics - Surgery Type

Class	**# of Patients**
Low Risk	9,559
High Risk	1,469
Surgery Type	**# of Patients**
Surgery A	95
Sugery B	3,005
Surgery C	487
Surgery D	25
Surgery E	95
Surgery F	4
Surgery G	1800
Surgery H	358
Surgery I	32
Surgery J	1,401
Surgery K	40
Surgery L	249
Surgery M	1,012
Surgery N	63
Surgery O	509
Surgery P	20
Surgery Q	62
Surgery R	7
Surgery S	11
Surgery T	124
Surgery U	1,378
Surgery V	69
Surgery W	182

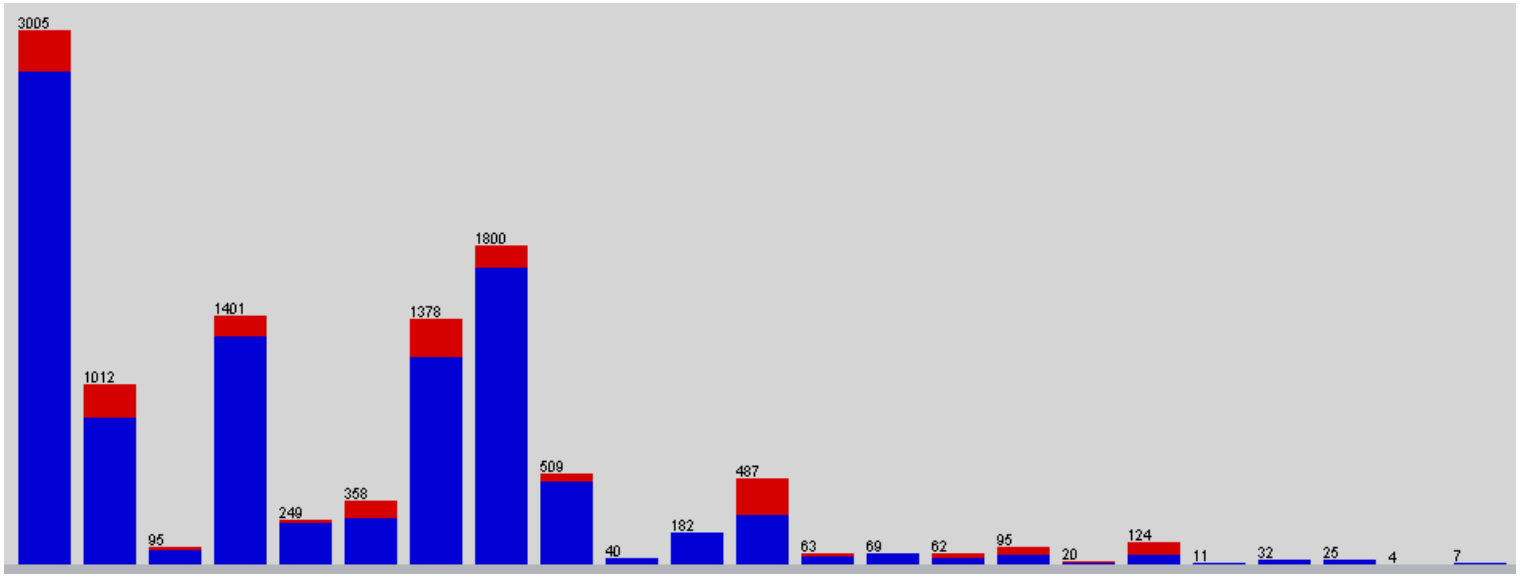

Figure 4.1: Surgery Type Distribution among Low Risk and High Risk Patients

Table 4.2: Low Risk and High Risk Patients Data Characteristics - Age, Gender, Race, ASA Physical Status, BMI

Age	Min Val. = 1, Max Val = 90, Mean = 60.081, St. Dev. = 14.52, Missing Val. = 1, Distinct Val. = 728, Unique Val. = 35
Gender	
Male	5,561
Female	5,467
Race	
Caucasian	9,322
African American	1,144
Other	393
ASA Physical Status	
I-II	5,315
III	5,255
IV-VI	456
BMI	Min Val. = 12.15, Max Val. = 92.59, Mean=30.06, St. Dev. 7.179, Missing = 1,222, Unique Val. = 819, Distinct Val. = 26

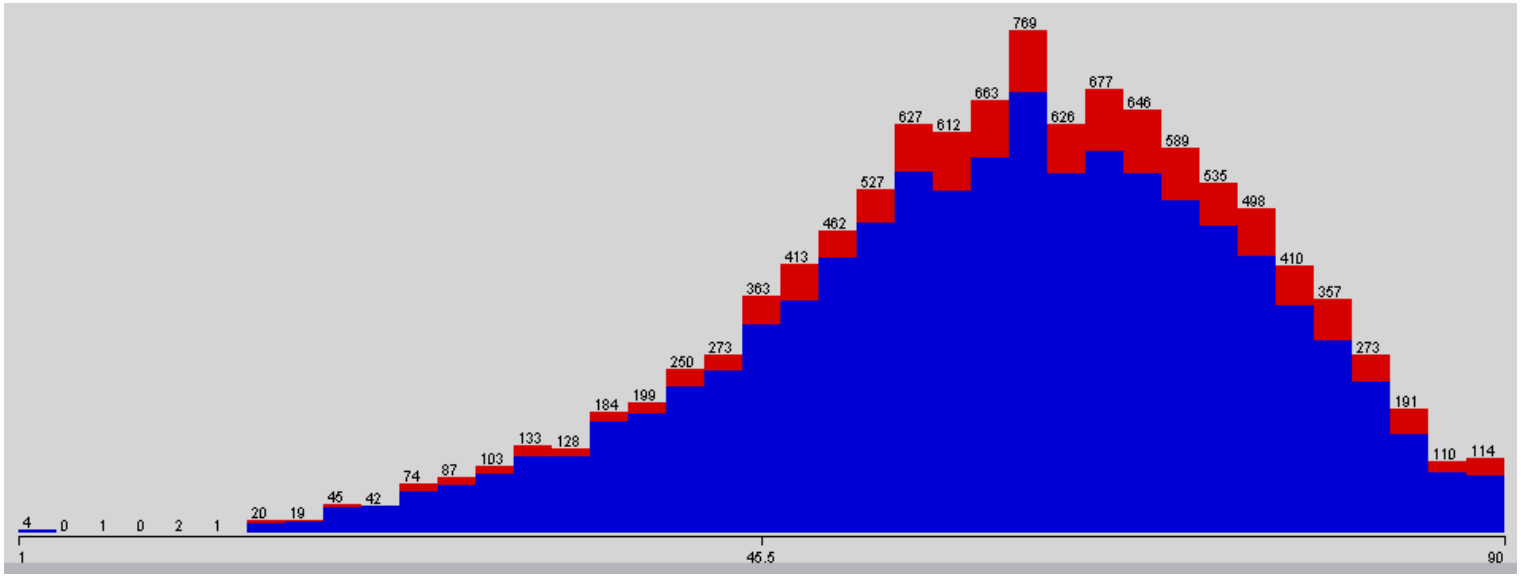

Figure 4.2: Age Distribution among Low Risk and High Risk Patients

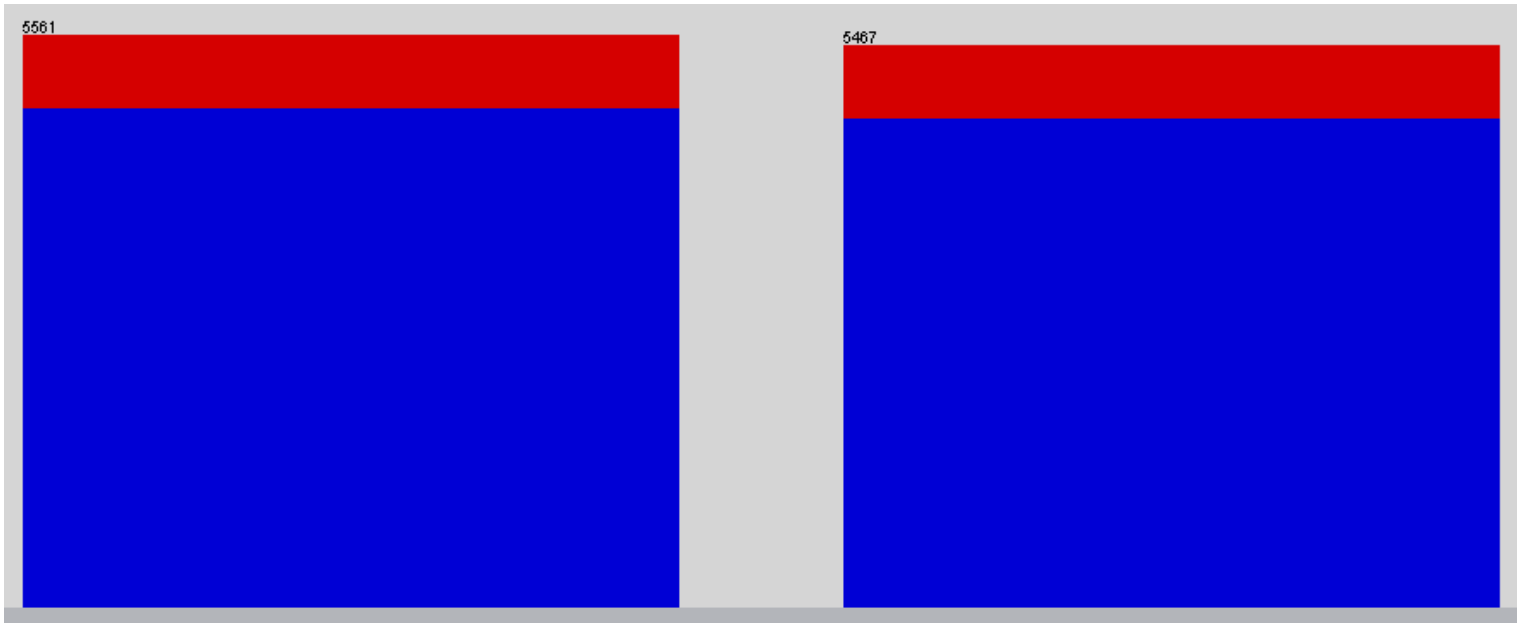

Figure 4.3: Gender Distribution among Low Risk and High Risk Patients

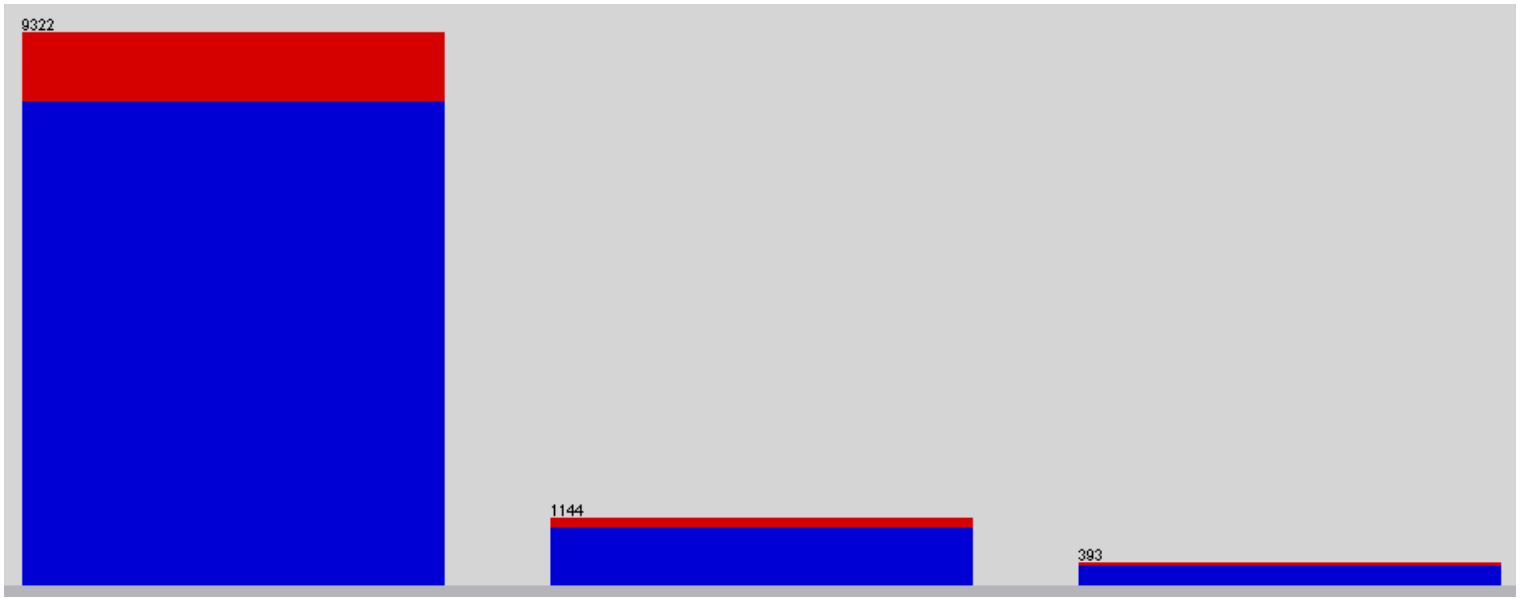

Figure 4.4: Race Distribution among Low Risk and High Risk Patients

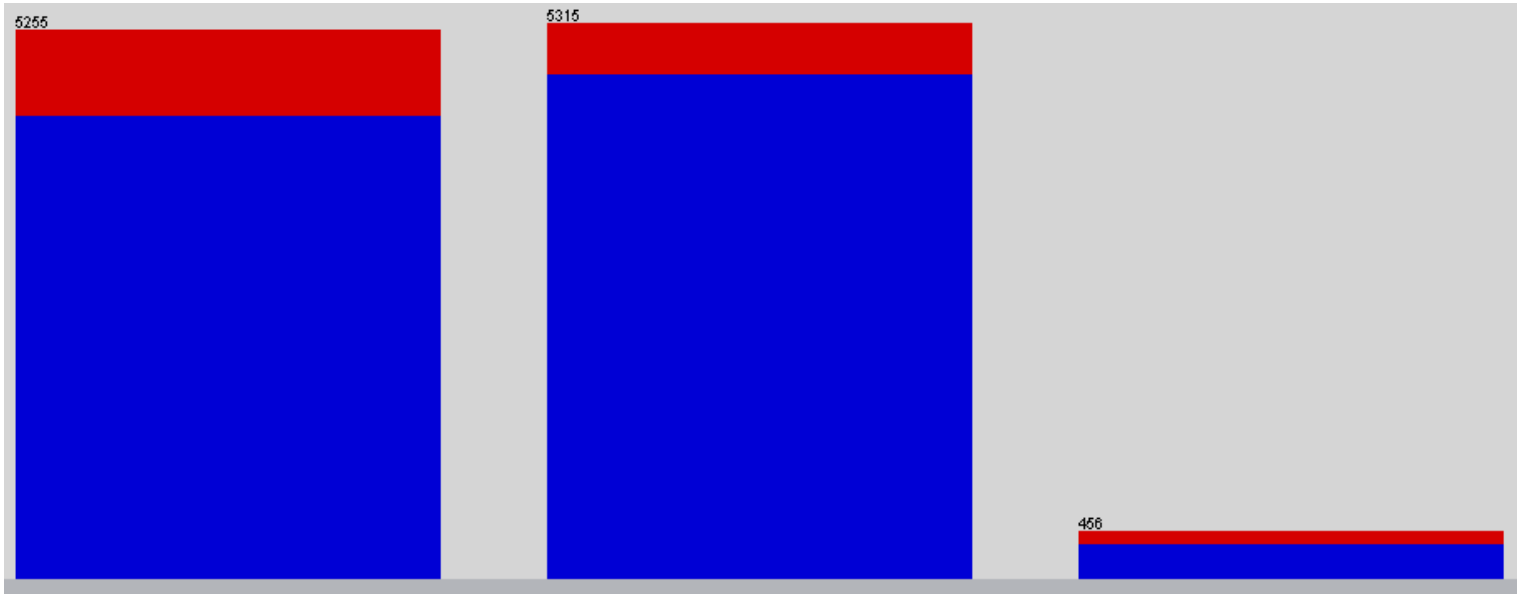

Figure 4.5: ASA Physical Status Distribution among Low Risk and High Risk Patients

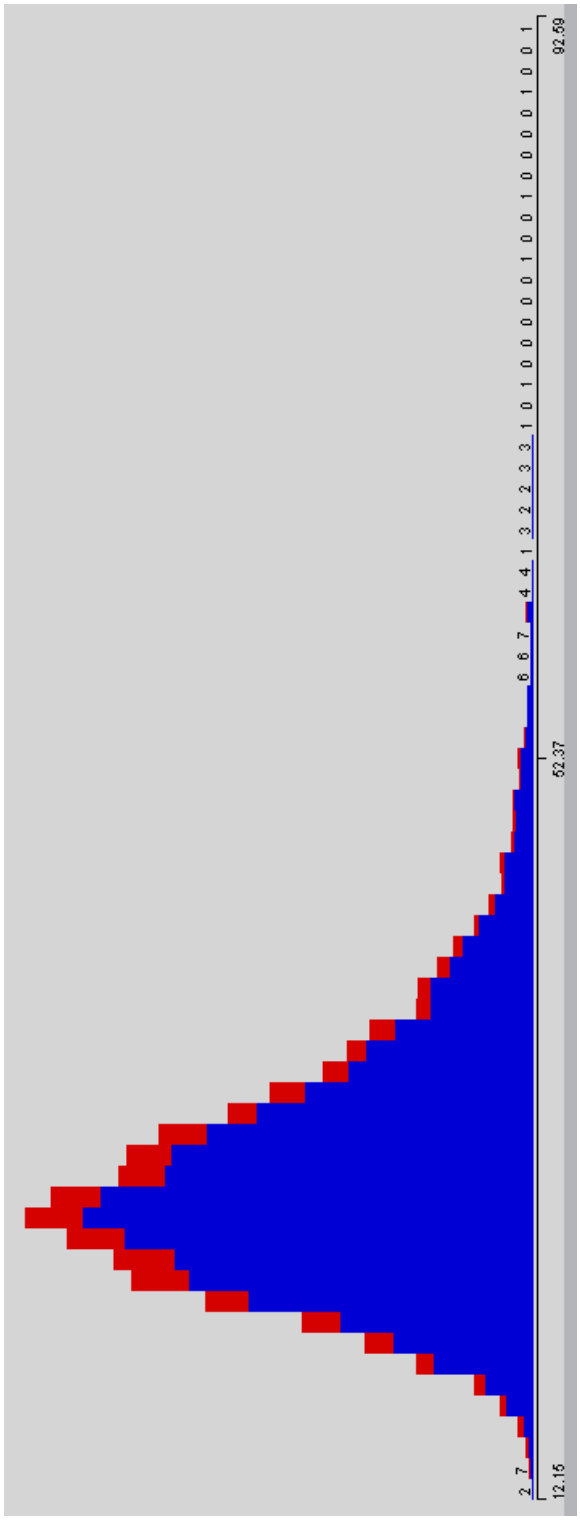

Figure 4.6: BMI Distribution among Low Risk and High Risk Patients

Table 4.3: Low Risk and High Risk Patients Data Characteristics - Underlying Health Conditions

Baseline Cancer	
No	7,930
Yes	3,098
Baseline CVD	
Yes	6,229
No	4,799
Baseline Dementia	
No	10,906
Yes	122
Baseline Diabetes	
No	9,451
Yes	1,577
Baseline Digestive	
Yes	2,472
No	8.556
Baseline Osteoart	
Yes	4,605
No	6,423
Baseline Psych	
No	10,005
Yes	1,023
Baseline Pulmonary	
No	9,793
Yes	1,235

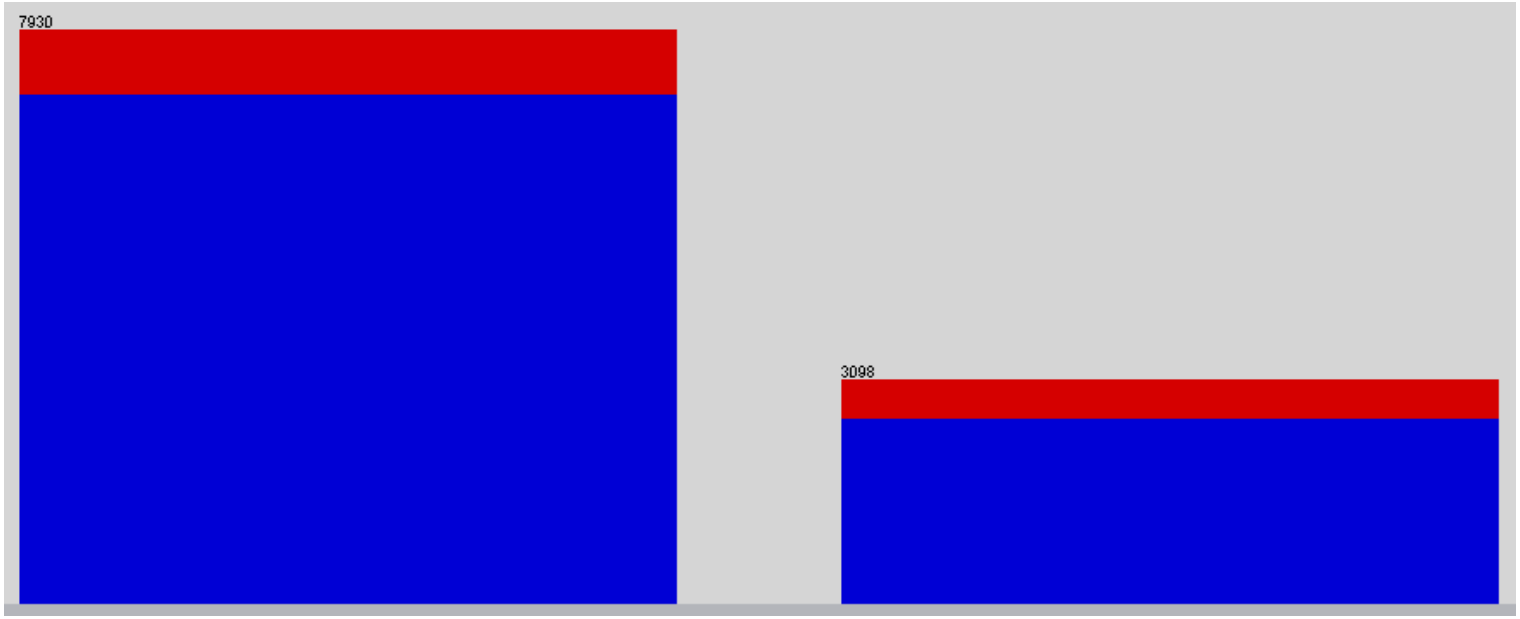

Figure 4.7: Baseline Cancer Distribution among Low Risk and High Risk Patients

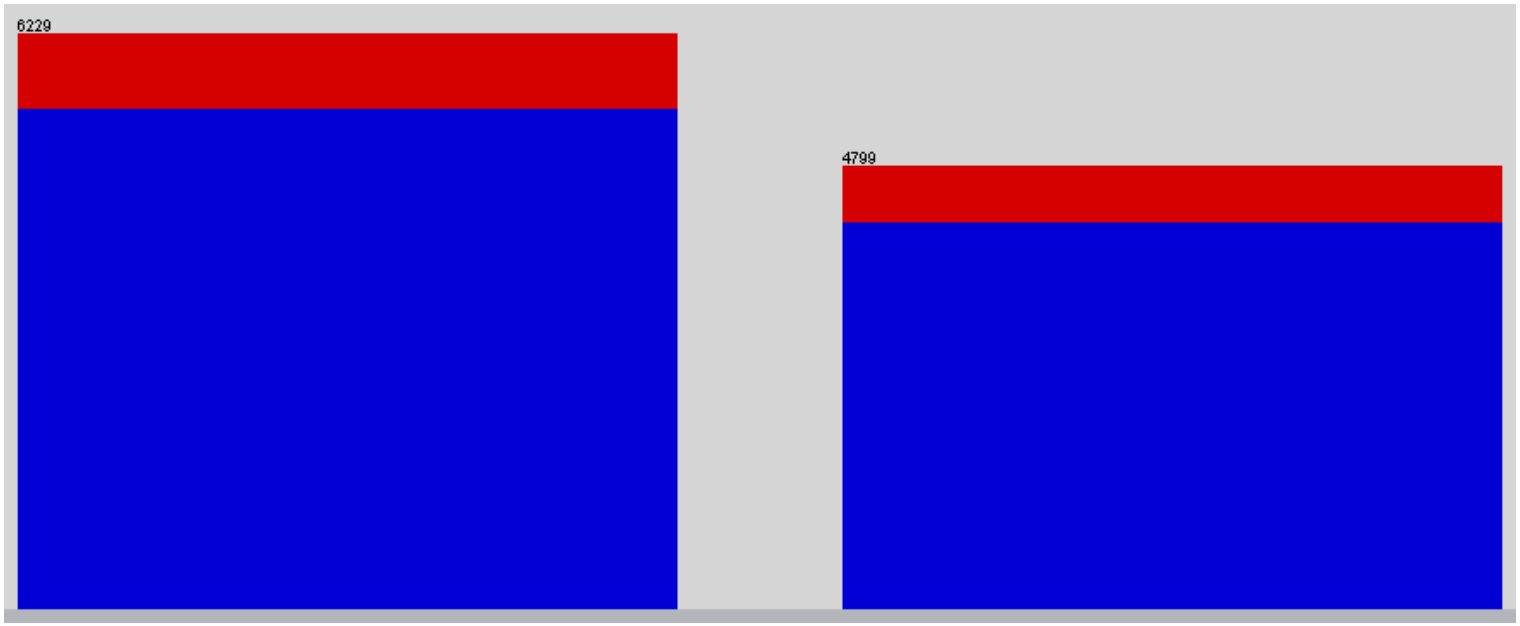

Figure 4.8: Baseline CVD Distribution among Low Risk and High Risk Patients

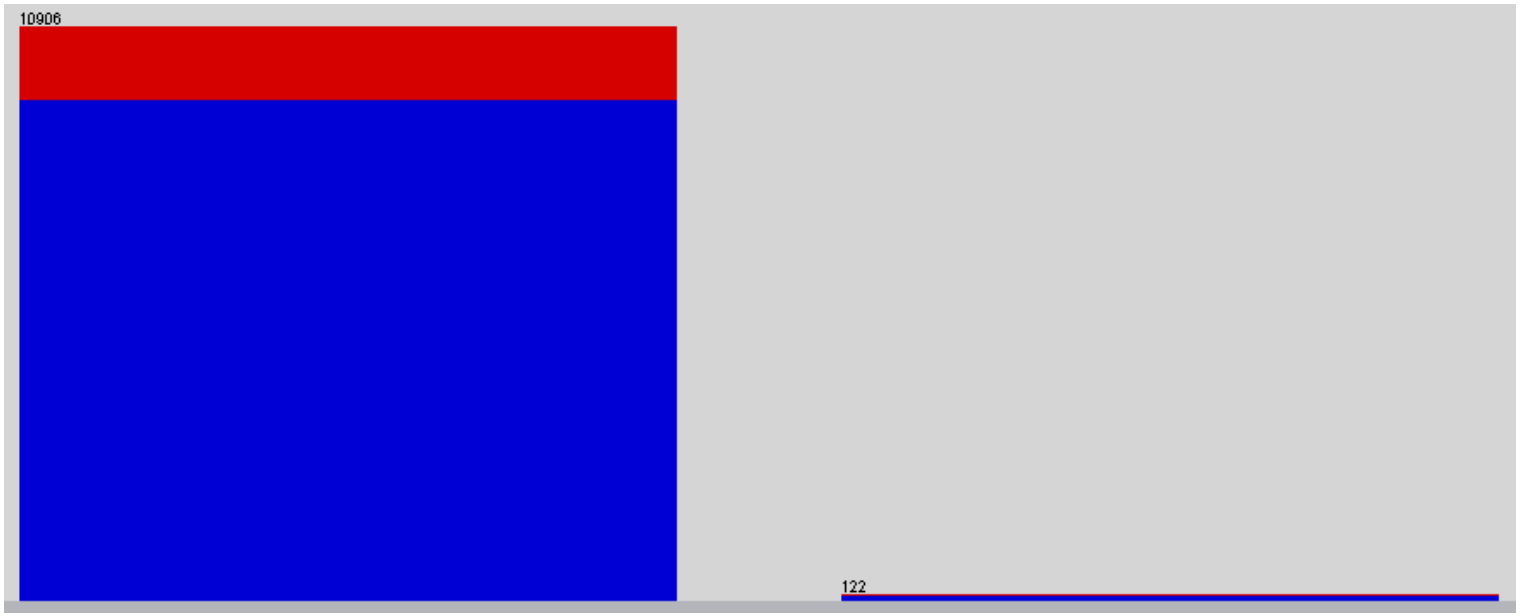

Figure 4.9: Baseline Dementia Distribution among Low Risk and High Risk Patients

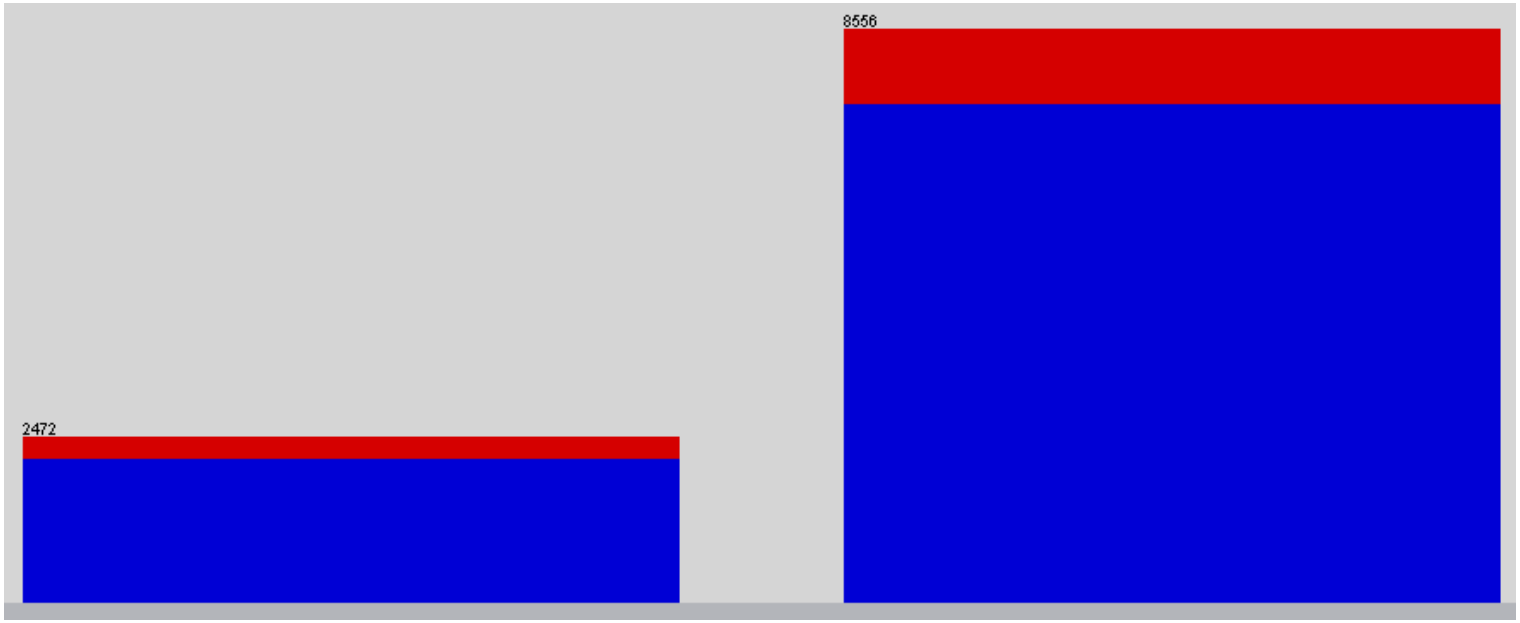

Figure 4.10: Baseline Digestive Disorder Distribution among Low Risk and High Risk Patients

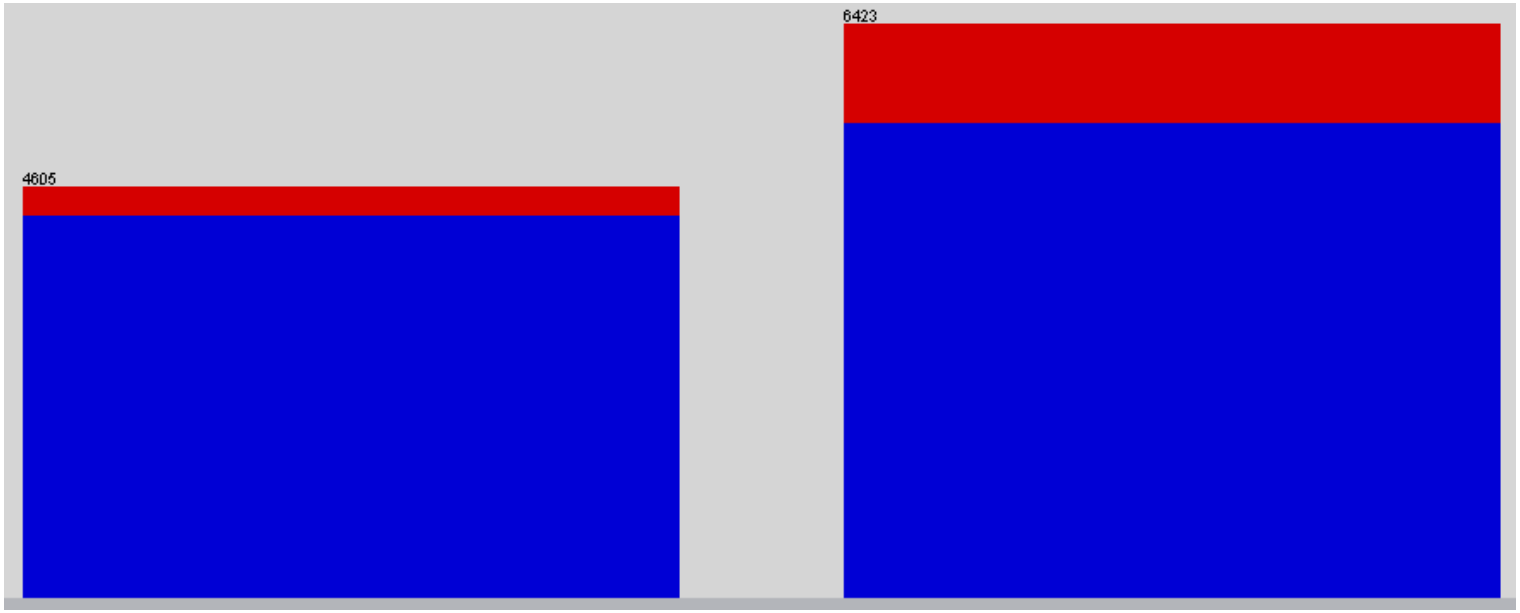

Figure 4.11: Baseline Osteoarthritis Distribution among Low Risk and High Risk Patients

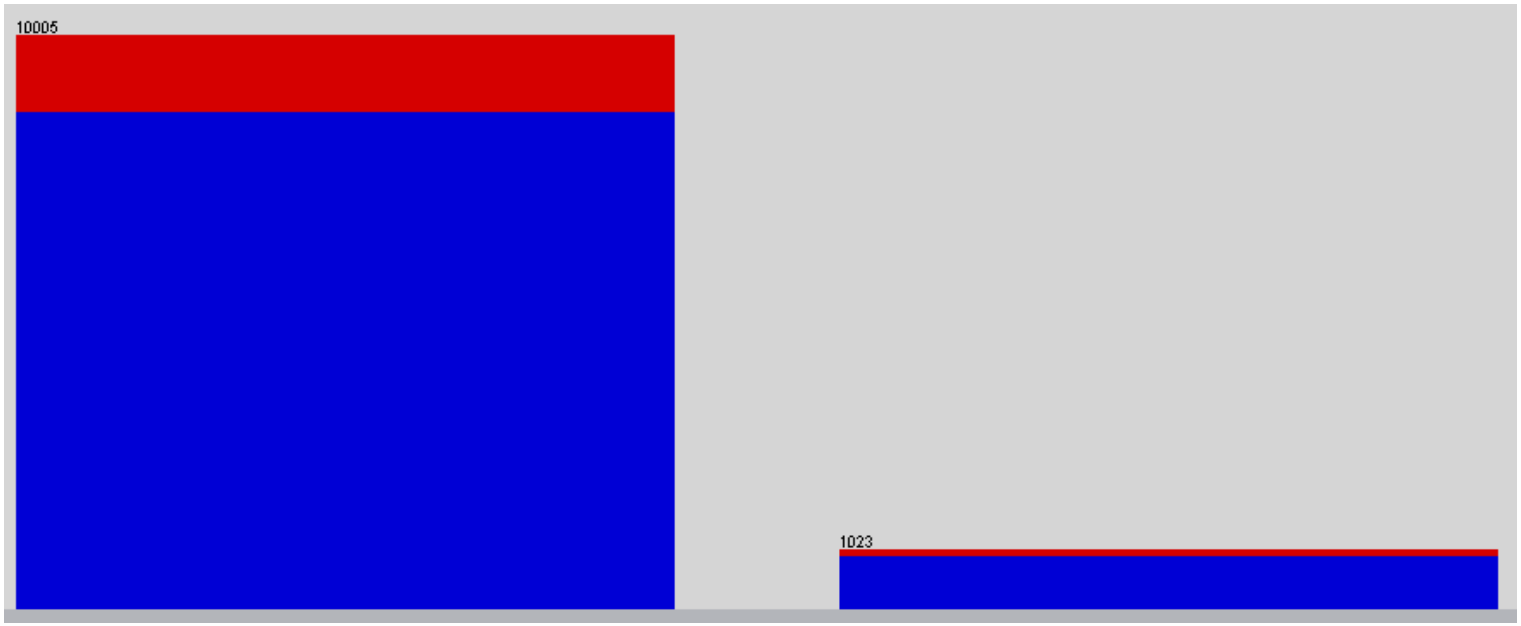

Figure 4.12: Baseline Psychiatric Disorder Distribution among Low Risk and High Risk Patients

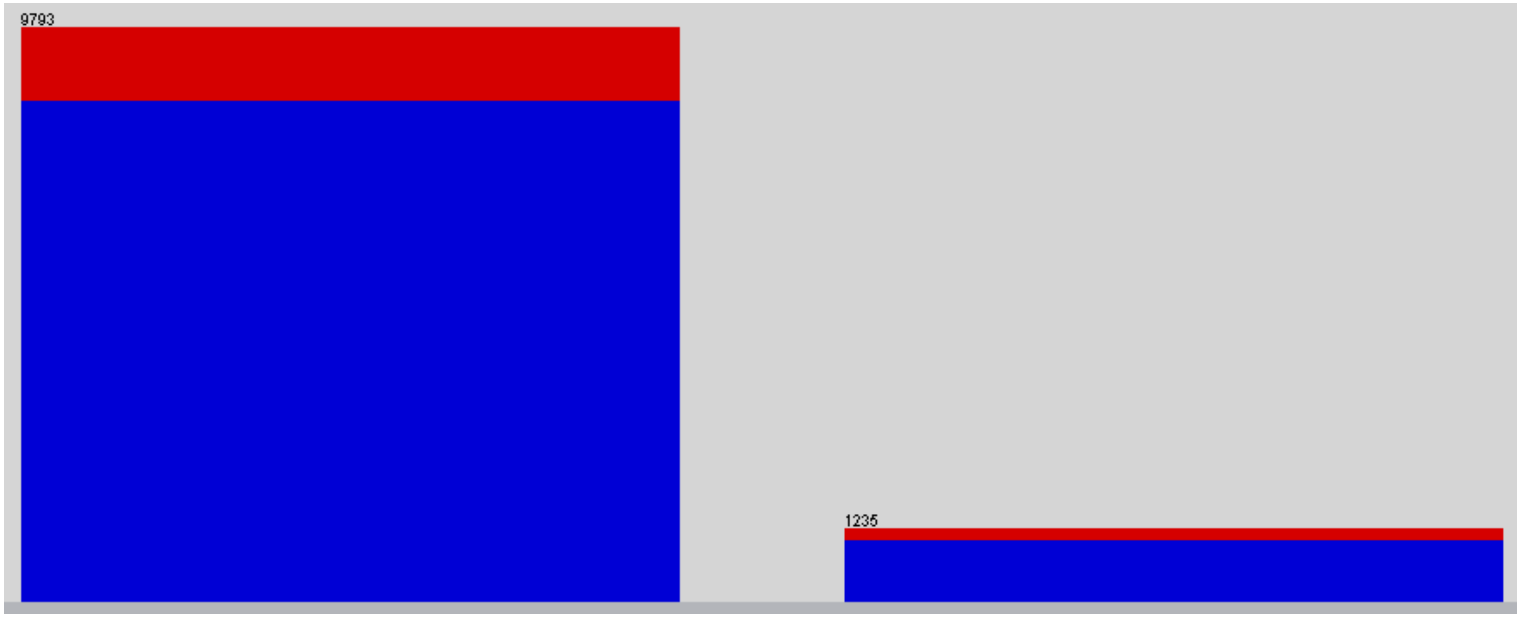

Figure 4.13: Baseline Pulmonary Disease Distribution among Low Risk and High Risk Patients

Table 4.4: Low Risk and High Risk Patients Data Characteristics - Baseline Charlson Index, Overall Incidence of 30-day Mortality for Each Surgery, and Overall Incidence of In-hospital Complications for Each Surgery

Baseline Charlson	Min Val. = 0, Max Val. = 13, Mean = 1.203, St. Dev. = 2.034, Missing Val. = 0, Distinct Val.= 14, Unique Val. = 1
ccsMort30rate	Min Val. = 0, Max Val. = 0.017. Mean = 0.005, St. Dev= 0.004, Missing Val. = 0, Distinct Val. = 21, Unique Val. = 0
ccsComplicationRate	Min Val. = 0.016, Max Val. = 0.466, Mean = 0.128, St. Dev. = 0.07, Missing Val. = 0, Distinct Val. = 23, Unique Val. = 0

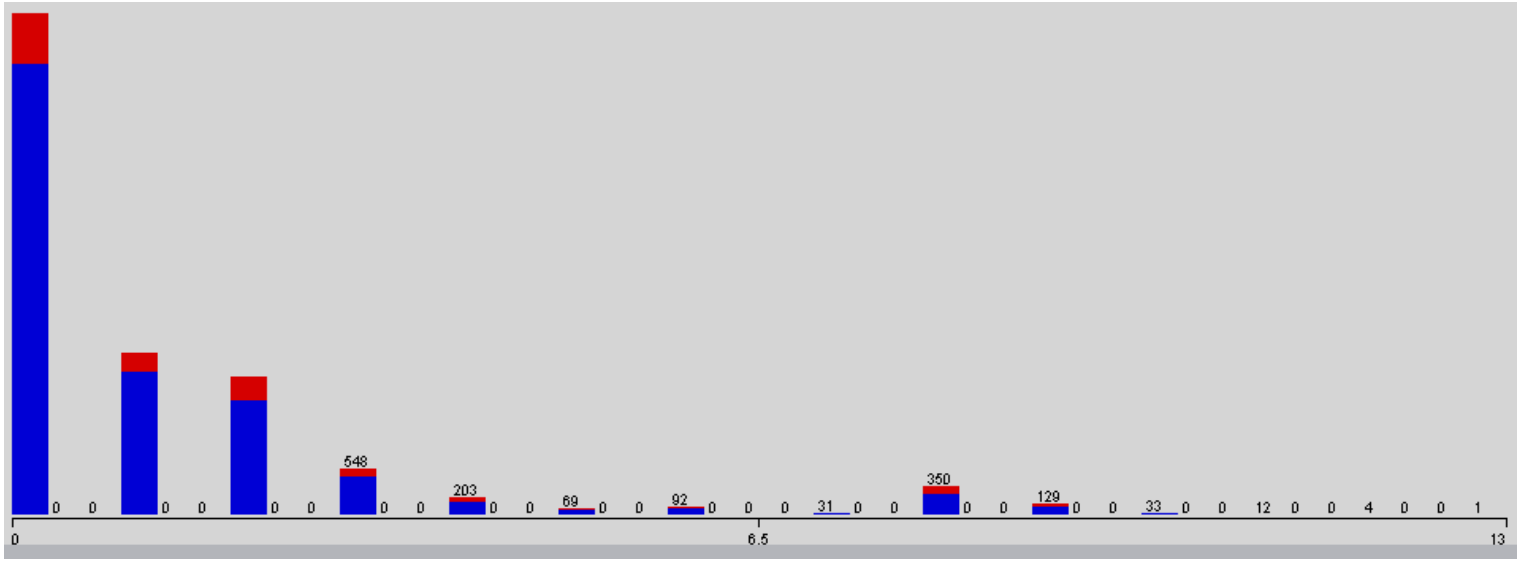

Figure 4.14: Baseline Charlson Index Distribution among Low Risk and High Risk Patients

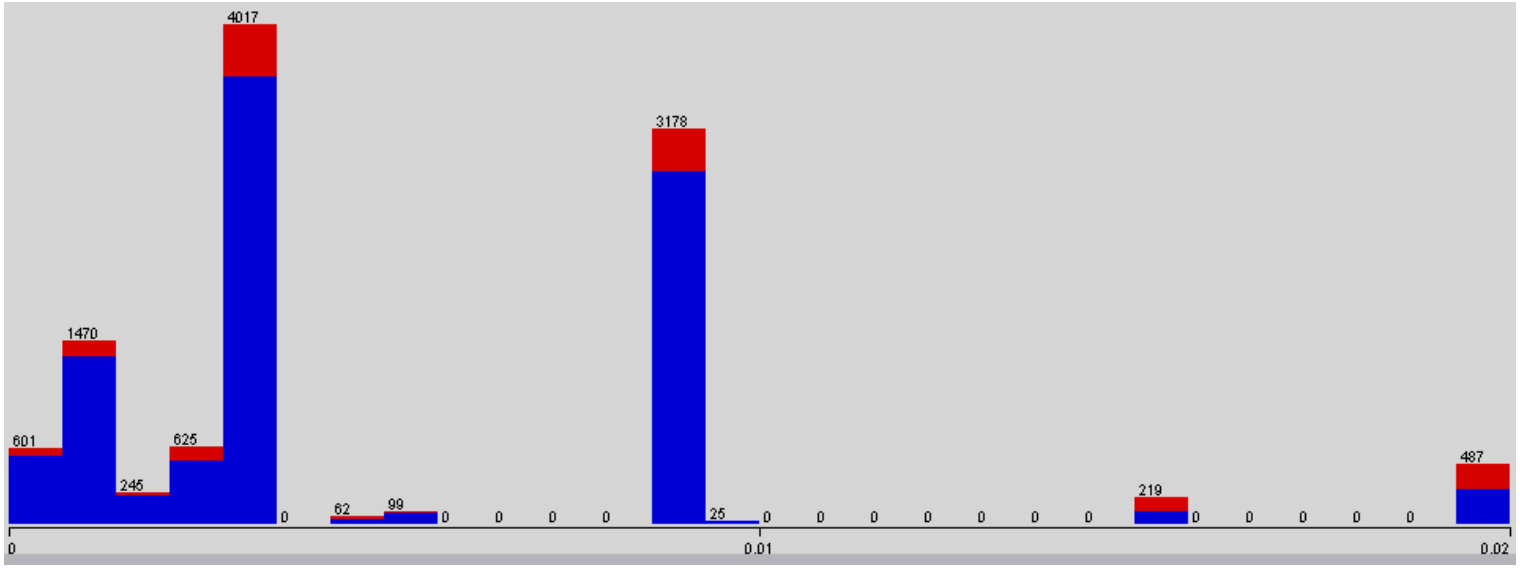

Figure 4.15: Overall Incidence of 30-day Mortality Distribution among Low Risk and High Risk Patients

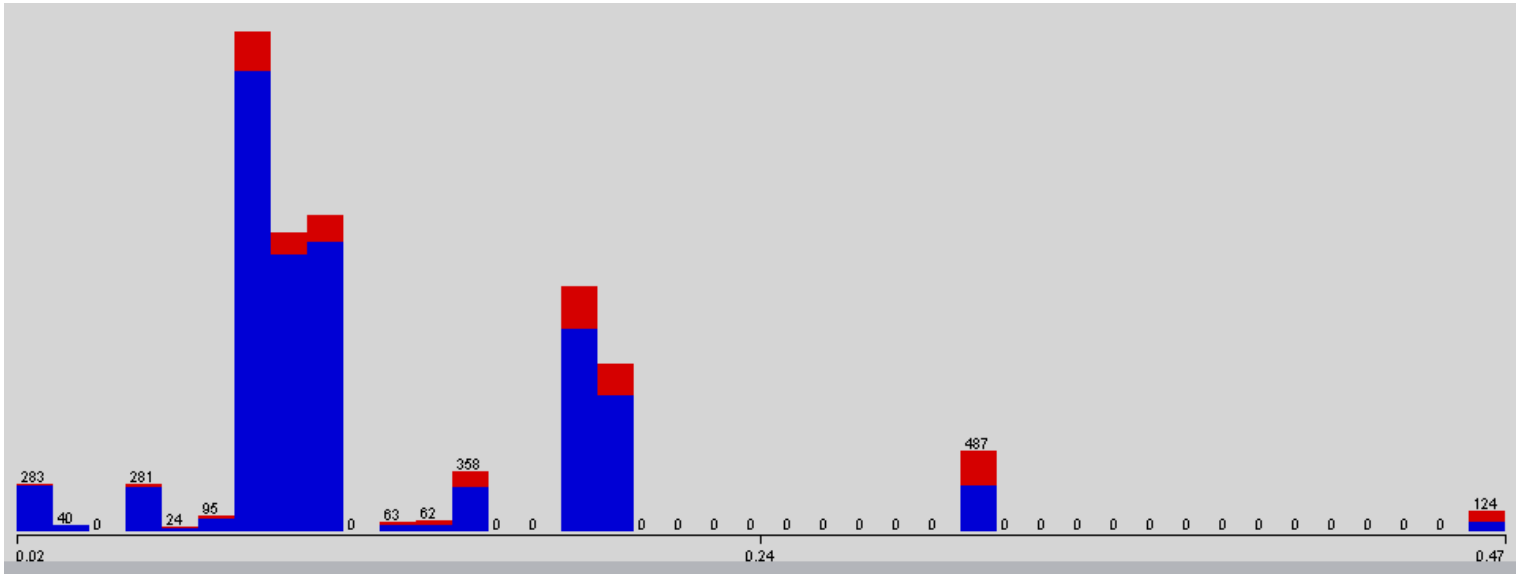

Figure 4.16: Overall Incidence of Complication Distribution among Low Risk and High Risk Patients

Table 4.5: Low Risk and High Risk Patients Data Characteristics - Hour, Day, Month, Moon Phase of Surgery, 30-Mortality of Patients, In-hospital Complication of Patients

Hour	Min Val. = 6.07, Max Val. = 19, Mean = 10.117, St. Dev. = 2.846, Missing Val. = 0, Distinct Val. = 724, Unique Val. = 57
Day of Week	
Monday	2,703
Tuesday	2,559
Wednesday	1,950
Thursday	1,952
Friday	1,864
Month	
January	937
February	884
March	943
April	974
May	946
June	1,026
July	770
August	1,073
September	1,109
October	953
November	817
December	596
Moon Phase	
First Quarter	2,820
Last Quarter	2,785
New Moon	2,639
Full Moon	2,784
mort30	
No	10,922
Yes	106
Complication	
No	9,589
Yes	1,439

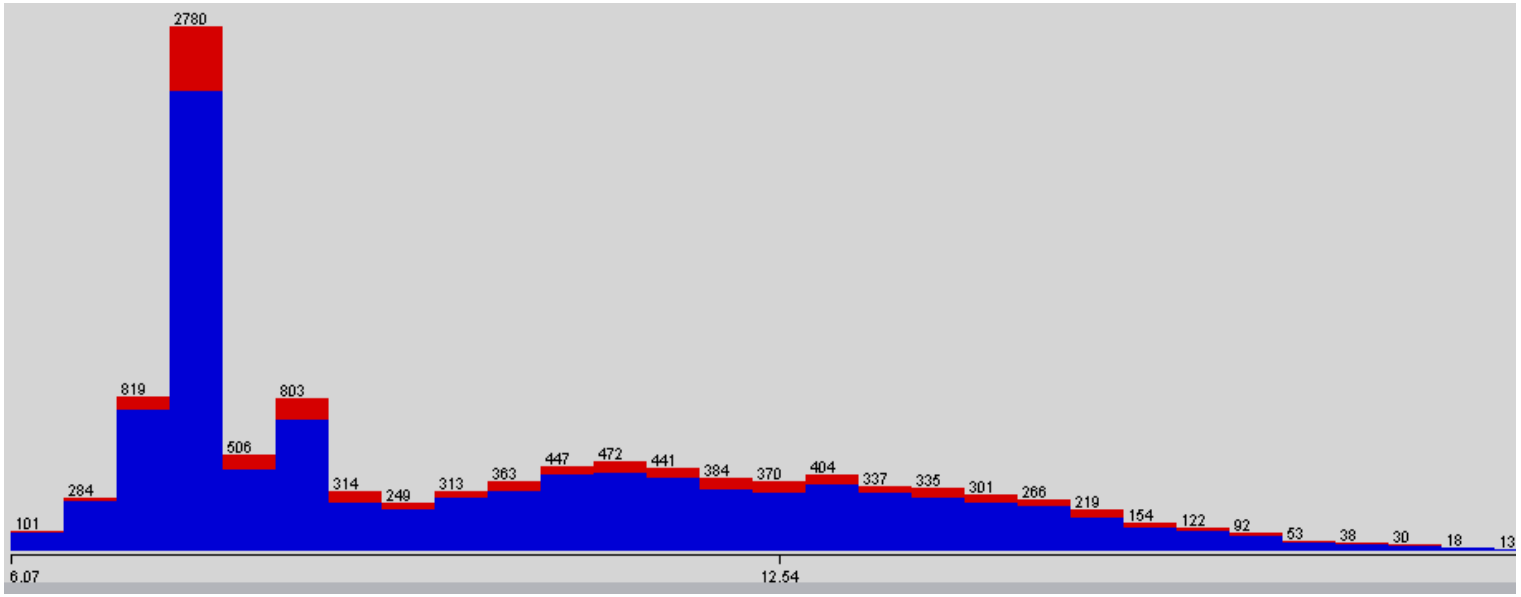

Figure 4.17: Hour of Surgery Distribution among Low Risk and High Risk Patients

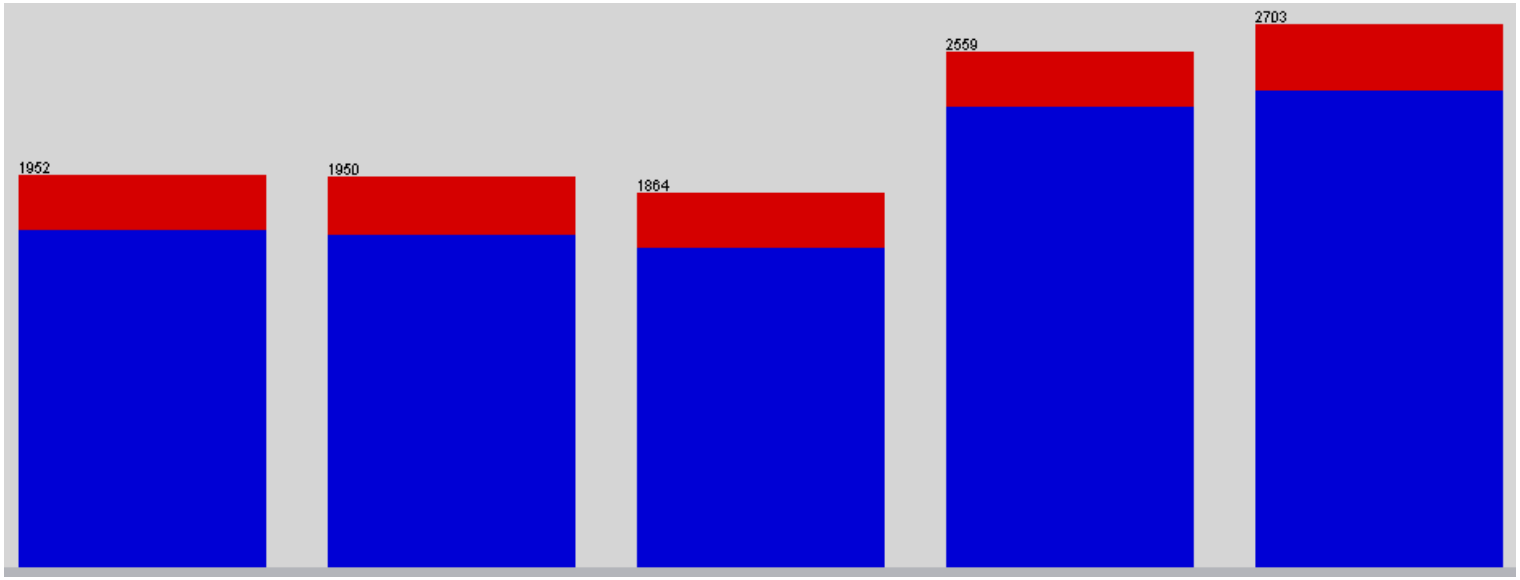

Figure 4.18: Day of Surgery Distribution among Low Risk and High Risk Patients

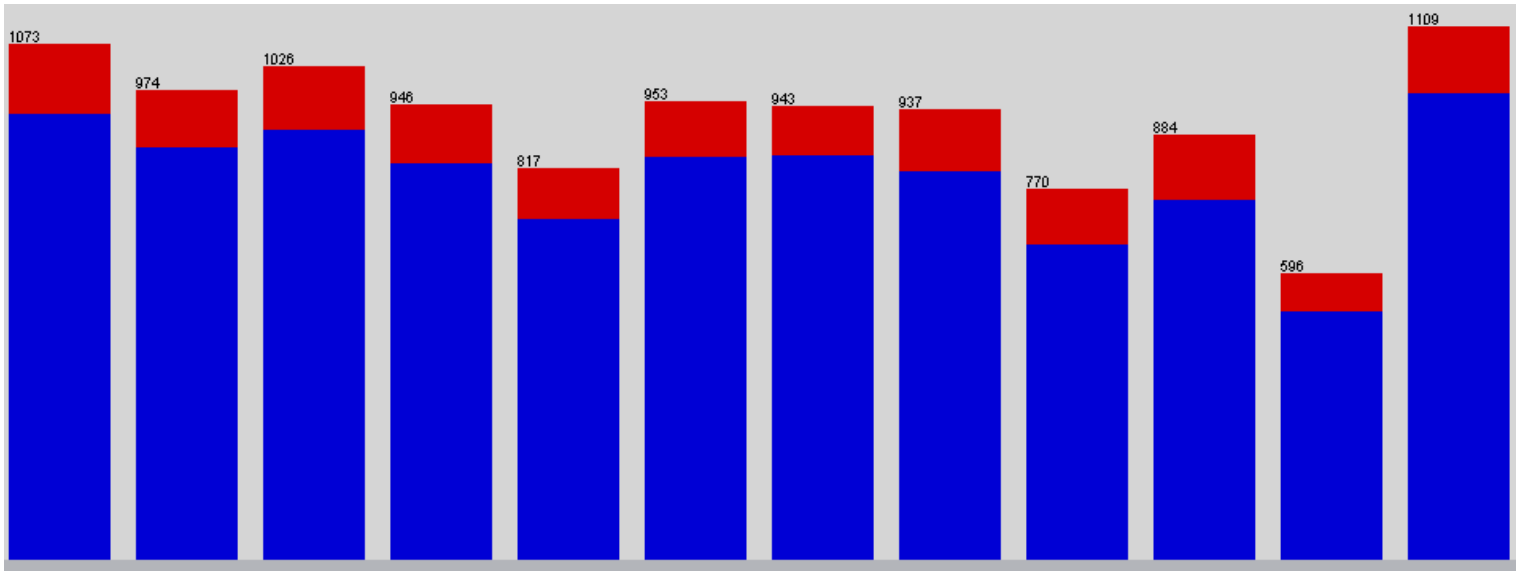

Figure 4.19: Month of Surgery Distribution among Low Risk and High Risk Patients

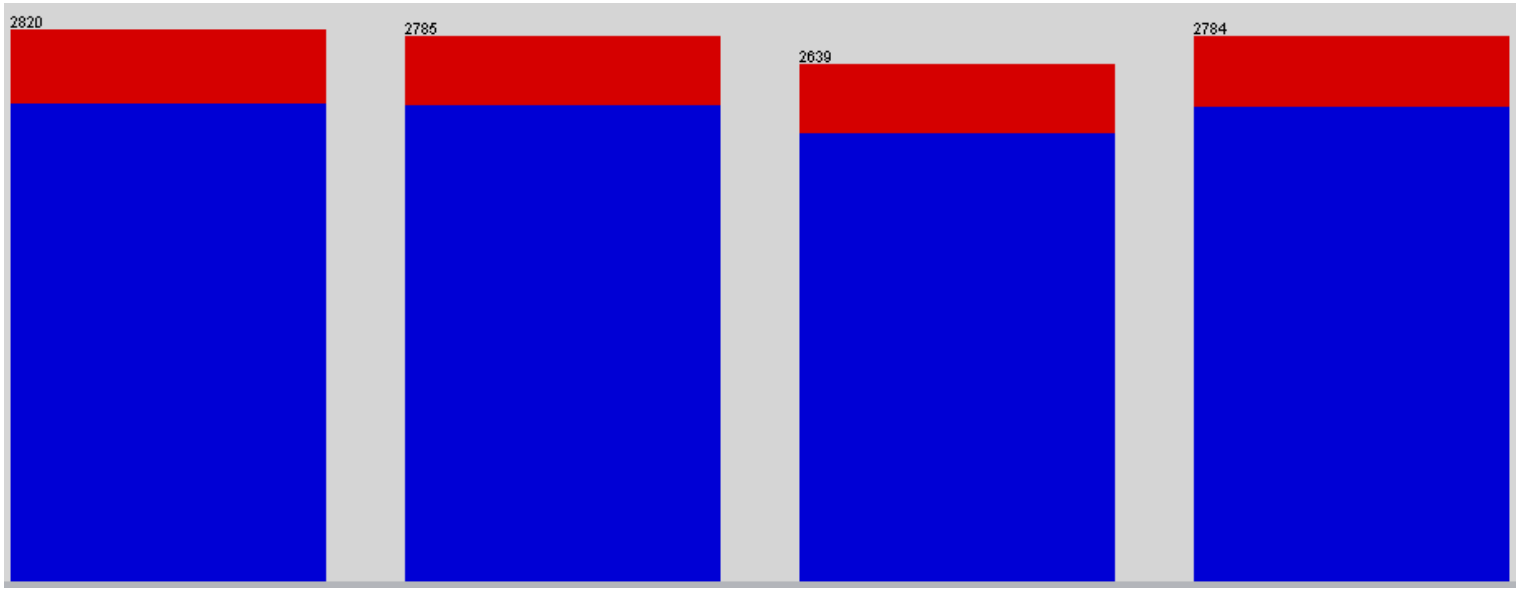

Figure 4.20: Moon Phase of Surgery Distribution among Low Risk and High Risk Patients

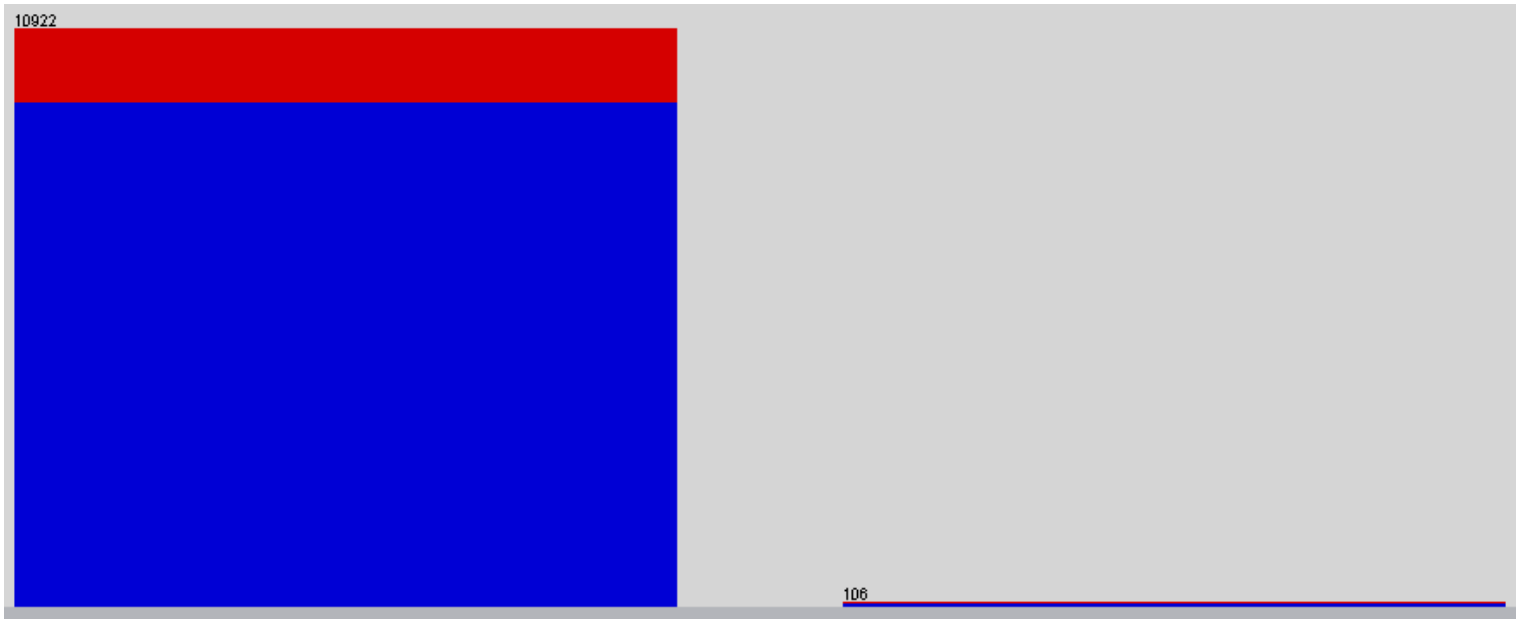

Figure 4.21: 30-day Patient Mortality Distribution among Low Risk and High Risk Patients

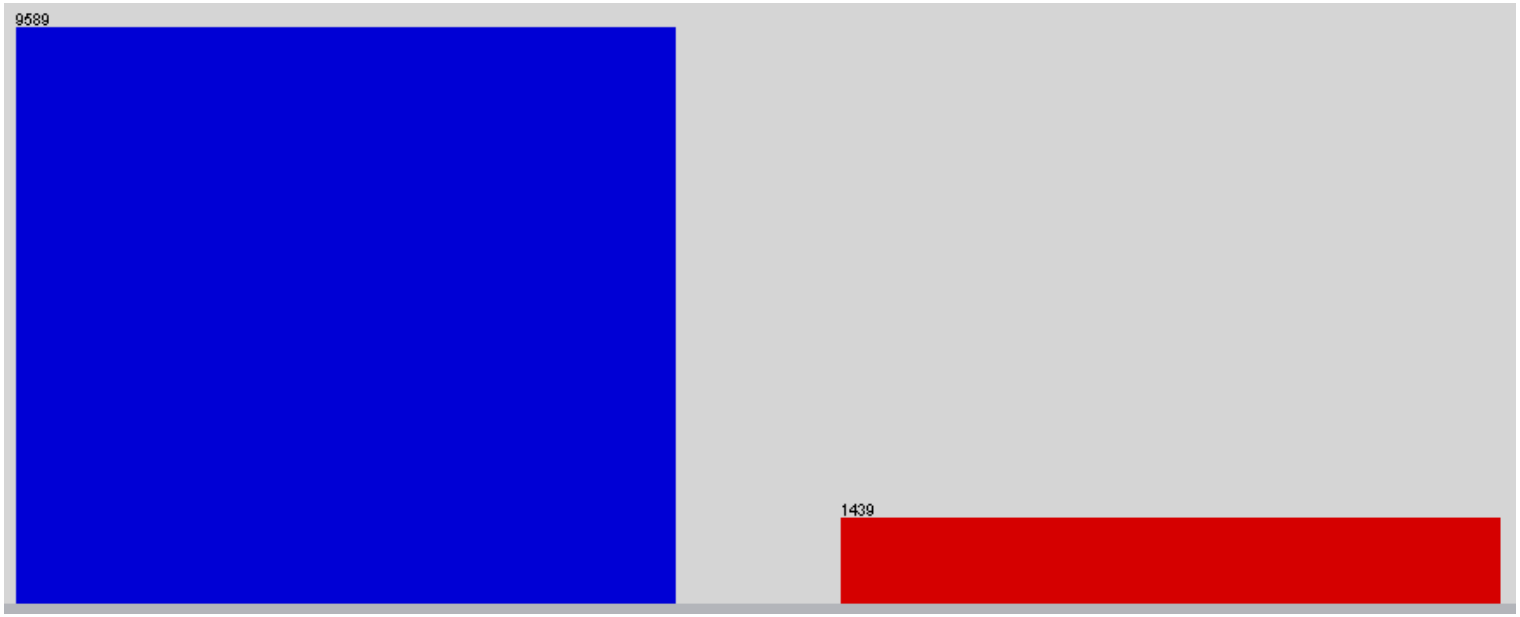

Figure 4.22: Patient Complication Distribution among Low Risk and High Risk Patients

4.1.2 Prediction of Low and High Risk Patients

To predict low risk and high risk patients in "Surgery Timing LH" dataset, we use 10×10-folding cross-validation experiments on nine commonly used and well-known classification methods, including Random Forest, Decision Trees, Nearest Neighbor, Stochastic Gradient Descent, Logistic Regression, Naïve Bayes, Bayes Network, Neural Networks, and Support Vector Machines [13]. Surgery Timing LH dataset is randomly partitioned into ten approximately equal parts; one of these subsets is designated as "test set", a model is built on the remaining nine subsets which form the "training dataset", and then tested by predicting the classes of patients in the test set using a classification method. This procedure is repeated 10 times, always taking another one of the ten parts in the role of the test set (re-randomizing the patients into 10 new subsets and repeat the procedure 9 additional times) for a total of 100 tests for each of the nine classification methods. Tables 4.6-4.14 show the average accuracy, proportion of correctly classified low risk patients, proportion of correctly classified high risk patients as well as average precision, recall, F-measure (weighted mean of the precision and recall), and area under Receiver Operating Characteristic (ROC) curve for Random Forest, Decision Trees, Nearest Neighbor, Stochastic Gradient Descent, Logistic Regression, Naïve Bayes, Bayes Network, Neural Networks, and Sequential Minimal Optimization, respectively.

Table 4.6: Cross-Validation of Low Risk and High Risk Patients Using Random Forest

Random Forest	**Average Cross-Validation Results**
Training Instances	9925.2
Testing Instance	1102.8
Number Correct	960.4
Number Incorrect	142.4
Percent Correct	87.1%
Percent Incorrect	12.9%
Mean Absolute Error	0.198
Area Under ROC	0.722
F-Measure	0.929
True Positive Rate	0.981
Number of True Positives	940.5
False Positive Rate	0.862
Number of False Positives	124.1
True Negative Rate	0.137
Number of True Negatives	19.8
False Negative Rate	0.019
Number of False Negatives	18.3
Weighted True Positive Rate	0.871
Weighted False Positive Rate	0.752
Weighted True Negative Rate	0.247
Weighted False Negative Rate	0.129
Weighted F-Measure	0.836
Weighted Area Under ROC	0.722

Table 4.7: Cross-Validation of Low Risk and High Risk Patients Using J48 Decision Tree

J48 Decision Tree	**Average Cross-Validation Results**
Training Instances	9925.2
Testing Instance	1102.8
Number Correct	963
Number Incorrect	139.8
Percent Correct	87.3%
Percent Incorrect	12.6%
Mean Absolute Error	0.194
Area Under ROC	0.643
F-Measure	0.931
True Positive Rate	0.976
Number of True Positives	936.1
False Positive Rate	0.812
Number of False Positives	116.9
True Negative Rate	0.187
Number of True Negatives	26.9
False Negative Rate	0.023
Number of False Negatives	22.8
Weighted True Positive Rate	0.873
Weighted False Positive Rate	0.709
Weighted True Negative Rate	0.290
Weighted False Negative Rate	0.127
Weighted F-Measure	0.845
Weighted Area Under ROC	0.643

Table 4.8: Cross-Validation of Low Risk and High Risk Patients Using k-Nearest Neighbor

k-Nearest Neighbor	**Average Cross-Validation Results**
Training Instances	9925.2
Testing Instance	1102.8
Number Correct	891.9
Number Incorrect	210.8
Percent Correct	80.8%
Percent Incorrect	19.1%
Mean Absolute Error	0.191
Area Under ROC	0.579
F-Measure	0.889
True Positive Rate	0.889
Number of True Positives	853.1
False Positive Rate	0.730
Number of False Positives	105.1
True Negative Rate	0.269
Number of True Negatives	38.7
False Negative Rate	0.110
Number of False Negatives	105.7
Weighted True Positive Rate	0.808
Weighted False Positive Rate	0.649
Weighted True Negative Rate	0.350
Weighted False Negative Rate	0.191
Weighted F-Measure	0.808
Weighted Area Under ROC	0.579

Table 4.9: Cross-Validation of Low Risk and High Risk Patients Using Stochastic Gradient Descent

Stochastic Gradient Descent	**Average Cross-Validation Results**
Training Instances	9925.2
Testing Instance	1102.8
Number Correct	960.2
Number Incorrect	142.6
Percent Correct	87.1%
Percent Incorrect	12.9%
Mean Absolute Error	0.129
Area Under ROC	0.521
F-Measure	0.930
True Positive Rate	0.994
Number of True Positives	953.4
False Positive Rate	0.952
Number of False Positives	137.1
True Negative Rate	0.047
Number of True Negatives	6.8
False Negative Rate	0.006
Number of False Negatives	5.5
Weighted True Positive Rate	0.871
Weighted False Positive Rate	0.829
Weighted True Negative Rate	0.171
Weighted False Negative Rate	0.129
Weighted F-Measure	0.820
Weighted Area Under ROC	0.521

Table 4.10: Cross-Validation of Low Risk and High Risk Patients Using Logistic Regression

Logistic Regression	**Average Cross-Validation Results**
Training Instances	9925.2
Testing Instance	1102.8
Number Correct	960.9
Number Incorrect	141.8
Percent Correct	87.1%
Percent Incorrect	12.8%
Mean Absolute Error	0.194
Area Under ROC	0.736
F-Measure	0.929
True Positive Rate	0.977
Number of True Positives	937.5
False Positive Rate	0.837
Number of False Positives	120.5
True Negative Rate	0.162
Number of True Negatives	23.3
False Negative Rate	0.022
Number of False Negatives	21.3
Weighted True Positive Rate	0.871
Weighted False Positive Rate	0.731
Weighted True Negative Rate	0.268
Weighted False Negative Rate	0.128
Weighted F-Measure	0.841
Weighted Area Under ROC	0.736

Table 4.11: Cross-Validation of Low Risk and High Risk Patients Using Naïve Bayes

Naïve Bayes	**Average Cross-Validation Results**
Training Instances	9925.2
Testing Instance	1102.8
Number Correct	924.6
Number Incorrect	178.2
Percent Correct	83.8%
Percent Incorrect	16.1%
Mean Absolute Error	0.181
Area Under ROC	0.723
F-Measure	0.907
True Positive Rate	0.906
Number of True Positives	863.3
False Positive Rate	0.615
Number of False Positives	88.6
True Negative Rate	0.384
Number of True Negatives	55.3
False Negative Rate	0.093
Number of False Negatives	89.5
Weighted True Positive Rate	0.838
Weighted False Positive Rate	0.547
Weighted True Negative Rate	0.452
Weighted False Negative Rate	0.161
Weighted F-Measure	0.838
Weighted Area Under ROC	0.722

Table 4.12: Cross-Validation of Low Risk and High Risk Patients Using Bayes Network

Bayes Network	**Average Cross-Validation Results**
Training Instances	9925.2
Testing Instance	1102.8
Number Correct	907.2
Number Incorrect	195.6
Percent Correct	82.2%
Percent Incorrect	17.7%
Mean Absolute Error	0.198
Area Under ROC	0.726
F-Measure	0.895
True Positive Rate	0.878
Number of True Positives	842.4
False Positive Rate	0.551
Number of False Positives	79.2
True Negative Rate	0.449
Number of True Negatives	64.7
False Negative Rate	0.121
Number of False Negatives	116.5
Weighted True Positive Rate	0.822
Weighted False Positive Rate	0.494
Weighted True Negative Rate	0.505
Weighted False Negative Rate	0.177
Weighted F-Measure	0.831
Weighted Area Under ROC	0.726

Table 4.13: Cross-Validation of Low Risk and High Risk Patients Using Multi-layer Perceptron

Multi-layer Perceptron	**Average Cross-Validation Results**
Training Instances	9925.2
Testing Instance	1102.8
Number Correct	926.9
Number Incorrect	175.8
Percent Correct	84.1%
Percent Incorrect	15.9%
Mean Absolute Error	0.163
Area Under ROC	0.665
F-Measure	0.911
True Positive Rate	0.931
Number of True Positives	892.9
False Positive Rate	0.764
Number of False Positives	109.9
True Negative Rate	0.236
Number of True Negatives	33.6
False Negative Rate	0.068
Number of False Negatives	65.9
Weighted True Positive Rate	0.841
Weighted False Positive Rate	0.673
Weighted True Negative Rate	0.326
Weighted False Negative Rate	0.159
Weighted F-Measure	0.827
Weighted Area Under ROC	0.655

Table 4.14: Cross-Validation of Low Risk and High Risk Patients Using Support Vector Machines

Support Vector Machines	**Average Cross-Validation Results**
Training Instances	9925.2
Testing Instance	1102.8
Number Correct	960.2
Number Incorrect	142.6
Percent Correct	87.1%
Percent Incorrect	12.9%
Mean Absolute Error	0.129
Area Under ROC	0.521
F Measure	0.930
True Positive Rate	0.994
Number of True Positives	953.4
False Positive Rate	0.953
Number of False Positives	137.1
True Negative Rate	0.047
Number of True Negatives	6.8
False Negative Rate	0.006
Number of False Negatives	5.5
Weighted True Positive Rate	0.871
Weighted False Positive Rate	0.829
Weighted True Negative Rate	0.171
Weighted False Negative Rate	0.129
Weighted F Measure	0.820
Weighted Area Under ROC	0.521

The average of 10×10-folding cross validation results for all nine classification methods are summarized in Table 4.15. The overall average accuracy of nine classification methods is 85.20%. Overall, the performance of the nine methods is validated by high values of prediction metrics: precision value of 0.89, recall value of 0.95, F-measure value of 0.92. Although not as significant, we observe that the overall average of value of areas under ROC curves is 0.65.

As can be seen from Table 4.15, all classification methods have comparable accuracy, precision, recall, F-measure and area under ROC curve. We also note that J48 Decision Tree method provides the combination of best accuracy, precision, recall, and F-value and its corresponding area under ROC curve is better than or comparable to those corresponding to the other classification methods.

Table 4.15: Average Cross-Validation Results for Nine Classfication Methods - Mortality RSI LH Data

Classification Method	**Accuracy**	**Precision**	**Recall**	**F-Measure**	**Area under ROC Curve**
Random Forest	87.09%	0.88	0.98	0.93	0.72
J48 Decision Tree	87.32%	0.89	0.98	0.93	0.64
Nearest Neighbor	80.86%	0.89	0.89	0.89	0.58
Stochastic Gradient Descent	87.08%	0.87	0.99	0.93	0.52
Logistic Regression	87.13%	0.89	0.98	0.93	0.74
Naïve Bayes	83.84%	0.91	0.91	0.91	0.72
Bayes Network	82.25%	0.91	0.88	0.90	0.73
Multilayer Perceptron	84.10%	0.89	0.93	0.91	0.67
Support Vector Machines	87.10%	0.87	0.99	0.93	0.52
Average	85.20%	0.89	0.95	0.92	0.65

4.1.3 Combinatorial Patterns of Low and High Risk Patients

Table 4.15 shows that J48 Decision Tree method provides us with best combination of average cross-validation accuracy, precision, recall, and F-measure. Because J48 Decision Tree method can be used to identify explicit/explainable patterns that can accurately predict low risk and high risk patients in Surgery Timing LH dataset, we apply the method on entire dataset to identify combinatorial patterns corresponding to low risk and high risk surgery patients stratified based on their *Mortality RSI* values.

The resulting J48 Decision Tree classification model consists of 83 low risk mortality patterns and 135 high risk mortality patterns presented in Tables 4.16-4.20 and Tables 4.21-4.29, respectively. Average 10×10-folding cross-validation accuracy of the J48 classification model is 87.32%. Average precision is 0.89 precision, recall is 0.98, F-measure value is 0.93, and value of the area under ROC curve is 0.64.

Table 4.16: Low Risk Mortality Patterns $L1$-$L15$

Pattern	Pattern Description
L1	$V17 \leq 0.031$ and $V16 \leq 0$ and $V23 = No$ and $V15 \leq 3$
L2	$V4 = Caucasian$ and $V15 \leq 1$ and $V3 = Female$ and $V8 = Yes$ and $V17 > 0.031$ and $V16 \leq 0$ and $V23 = No$
L3	$V4 = Other$ and $V15 \leq 1$ and $V3 = Female$ and $V8 = Yes$ and $V17 > 0.031$ and $V16 \leq 0$ and $V23 = No$
L4	$V6 \leq 30.56$ and $V8 = No$ and $V17 > 0.031$ and $V16 \leq 0$ and $V23 = No$
L5	$V12 = Yes$ and $V16 > 0$ and $V23 = No$ and $V15 \leq 3$
L6	$V16 \leq 0.00141$ and $V5 = III$ and $V9 = No$ and $V22 = No$ and $V12 = No$
L7	$V1 = B$ and $V16 > 0.00141$ and $V5 = III$ and $V9 = No$ and $V22 = No$ and $V12 = No$
L8	$V1 = M$ and $V16 > 0.0141$ and $V5 = III$ and $V9 = No$ and $V22 = No$ and $V12 = No$
L9	$V18 \leq 11.6$ and $V1 = A$ and $V16 > 0.0141$ and $V5 = III$ and $V9 = No$ and $V22 = No$ and $V12 = No$
L10	$V1 = J$ and $V16 > 0.0141$ and $V5 = III$ and $V9 = No$ and $V22 = No$ and $V12 = No$
L11	$V1 = L$ and $V16 > 0.0141$ and $V5 = III$ and $V9 = No$ and $V22 = No$ and $V12 = No$
L12	$V1 = H$ and $V16 > 0.0141$ and $V5 = III$ and $V9 = No$ and $V22 = No$ and $V12 = No$
L13	$V14 = No$ and $V15 \leq 1$ and $V1 = U$ and $V16 > 0.0141$ and $V5 = III$ and $V9 = No$ and $V22 = No$ and $V12 = No$
L14	$V15 > 0$ and $V14 = Yes$
L15	$V2 \leq 74.2$ and $V8 = Yes$ and $V10 = No$ and $V15 > 1$ and $V1 = U$ and $V16 > 0.0141$ and $V5 = III$ and $V9 = No$ and $V22 = No$ and $V12 = No$

Table 4.17: Low Risk Mortality Patterns $L16$-$L30$

Pattern	Pattern Description
L16	$V10 = Yes$ and $V15 > 1$ and $V1 = U$ and $V16 > 0.0141$ and $V5 = III$ and $V9 = No$ and $V22 = No$ and $V12 = No$
L17	$V2 \leq 61.1$ and $V1 = G$ and $V16 > 0.0141$ and $V5 = III$ and $V9 = No$ and $V22 = No$ and $V12 = No$
L18	$V18 \leq 14.2$ and $V13 = No$ and $V2 > 61.1$ and $V1 = G$ and $V16 > 0.0141$ and $V5 = III$ and $V9 = No$ and $V22 = No$ and $V12 = No$
L19	$V1 = O$ and $V16 > 0.0141$ and $V5 = III$ and $V9 = No$ and $V22 = No$ and $V12 = No$
L20	$V1 = K$ and $V16 > 0.0141$ and $V5 = III$ and $V9 = No$ and $V22 = No$ and $V12 = No$
L21	$V1 = W$ and $V16 > 0.0141$ and $V5 = III$ and $V9 = No$ and $V22 = No$ and $V12 = No$
L22	$V1 = C$ and $V16 > 0.0141$ and $V5 = III$ and $V9 = No$ and $V22 = No$ and $V12 = No$
L23	$V15 \leq 2$ and $V1 = N$ and $V16 > 0.0141$ and $V5 = III$ and $V9 = No$ and $V22 = No$ and $V12 = No$
L24	$V1 = V$ and $V16 > 0.0141$ and $V5 = III$ and $V9 = No$ and $V22 = No$ and $V12 = No$
L25	$V6 > 30.68$ and $V15 \leq 0$ and $V10 = No$
L26	$V10 = Yes$ and $V15 \leq 1$ and $V1 = Q$ and $V16 > 0.00141$ and $V5 = III$ and $V9 = No$ and $V22 = No$ and $V12 = No$
L27	$V1 = E$ and $V16 \leq 0.00141$ and $V5 = III$ and $V9 = No$ and $V22 = No$ and $V12 = No$
L28	$V1 = P$ and $V16 \leq 0.00141$ and $V5 = III$ and $V9 = No$ and $V22 = No$ and $V12 = No$
L29	$V1 = T$ and $V16 \leq 0.00141$ and $V5 = III$ and $V9 = No$ and $V22 = No$ and $V12 = No$
L30	$V1 = S$ and $V16 \leq 0.00141$ and $V5 = III$ and $V9 = No$ and $V22 = No$ and $V12 = No$

Table 4.18: Low Risk Mortality Patterns $L31$-$L45$

Pattern	Pattern Description
L31	$V1 = I$ and $V16 \leq 0.00141$ and $V5 = III$ and $V9 = No$ and $V22 = No$ and $V12 = No$
L32	$V1 = D$ and $V16 \leq 0.00141$ and $V5 = III$ and $V9 = No$ and $V22 = No$ and $V12 = No$
L33	$V1 = F$ and $V16 \leq 0.00141$ and $V5 = III$ and $V9 = No$ and $V22 = No$ and $V12 = No$
L34	$V5 = I - II$ and $V9 = No$ and $V22 = No$ and $V22 = No$ and $V12 = No$ and $V16 > 0$ and $V23 = No$ and $V15 \leq 3$
L35	$V17 \leq 0.084034$ and $V5 = IV - VI$ and $V9 = No$ and $V22 = No$ and $V12 = No$ and $V16 > 0$ and $V23 = No$ and $V15 \leq 3$
L36	$V2 > 65.6$ and $V1 = M$ and $V13 = No$ and $V10 = No$ and $V17 > 0.084034$ and $V5 = IV - VI$ and $V9 = No$ and $V22 = No$ and $V12 = No$ and $V16 > 0$ and $V23 = No$ and $V15 \leq 3$
L37	$V6 > 28.9$ and $V1 = J$ and $V13 = No$ and $V10 = No$ and $V17 > 0.084034$ and $V5 = IV - VI$ and $V9 = No$ and $V22 = No$ and $V12 = No$ and $V16 > 0$ and $V23 = No$ and $V15 \leq 3$
L38	$V14 = No$ and $V8 = Yes$ and $V15 \leq 1$ and $V1 = U$ and $V13 = No$ and $V10 = No$ and $V17 > 0.084034$ and $V5 = IV - VI$ and $V9 = No$ and $V22 = No$ and $V12 = No$ and $V16 > 0$ and $V23 = No$
L39	$V13 = Yes$ and $V10 = No$ and $V17 > 0.084034$ and $V5 = IV - VI$ and $V9 = No$ and $V22 = No$ and $V12 = No$ and $V16 > 0$ and $V23 = No$ and $V15 \leq 3$
L40	$V15 \leq 2$ and $V10 = Yes$ and $V17 > 0.084034$ and $V5 = IV - VI$ and $V9 = No$ and $V22 = No$ and $V12 = No$ and $V16 > 0$ and $V23 = No$
L41	$V2 > 73.9$ and $V15 > 2$ and $V10 = Yes$ and $V17 > 0.084034$ and $V5 = IV - VI$ and $V9 = No$ and $V22 = No$ and $V12 = No$ and $V16 > 0$ and $V23 = No$
L42	$V13 = Yes$ and $V9 = Yes$ and $V22 = No$ and $V12 = No$ and $V16 > 0$ and $V23 = No$ and $V15 \leq 3$
L43	$V12 = Yes$ and $V23 = Yes$ and $V15 \leq 3$
L44	$V18 \leq 7.4$ and $V10 = No$ and $V13 = No$ and $V5 = III$ and $V12 = No$ and $V23 = Yes$ and $V15 \leq 3$
L45	$V2 \leq 77.4$ and $V1 = B$ and $V18 > 7.4$ and $V10 = No$ and $V13 = No$ and $V5 = III$ and $V12 = No$ and $V23 = Yes$ and $V15 \leq 3$

Table 4.19: Low Risk Mortality Patterns $L46$-$L62$

Pattern	Pattern Description
L46	$V15 \leq 1$ and $V1 = M$ and $V18 > 7.4$ and $V10 = No$ and $V13 = No$ and $V5 = III$ and $V12 = No$ and $V23 = Yes$
L47	$V11 = Yes$ and $V1 = J$ and $V18 > 7.4$ AND $V10 = No$ and $V13 = No$ and $V5 = III$ and $V12 = No$ and $V23 = Yes$ and $V15 \leq 3$
L48	$V15 \leq 0$ and $V4 = Caucasian$ and $V11 = No$ and $V1 = J$ and $V18 > 7.4$ and $V10 = No$ and $V13 = No$ and $V5 = III$ and $V12 = No$ and $V23 = Yes$
L49	$V18 > 9.75$ and $V15 > 0$ and $V4 = Caucasian$ and $V11 = No$ and $V1 = J$ and $V18 > 7.4$ and $V10 = No$ and $V13 = No$ and $V5 = III$ and $V12 = No$ and $V23 = Yes$
L50	$V1 = L$ and $V18 > 7.4$ AND $V10 = No$ and $V13 = No$ and $V5 = III$ and $V12 = No$ and $V23 = Yes$ and $V15 \leq 3$
L51	$V11 = No$ and $V7 = Yes$ and $V1 = H$ and $V18 > 7.4$ and $V10 = No$ and $V13 = No$ and $V5 = III$ and $V12 = No$ and $V23 = Yes$ and $V15 \leq 3$
L52	$V18 \leq 7.55$ and $V1 = U$
L53	$V4 = Caucasian$ and $V11 = Yes$ and $V18 > 7.55$ and $V1 = U$
L54	$V4 = Other$ and $V11 = Yes$ and $V18 > 7.55$ and $V1 = U$
L55	$V4 = AfricanAmerican$ and $V11 = No$ and $V18 > 7.55$ and $V1 = U$
L56	$V2 \leq 64.4$ and $V1 = G$ and $V18 > 7.4$ and $V10 = No$ and $V13 = No$ and $V5 = III$ and $V12 = No$ and $V23 = Yes$ and $V15 \leq 3$
L57	$V1 = O$ and $V18 > 7.4$ and $V10 = No$ and $V13 = No$ and $V5 = III$ and $V12 = No$ and $V23 = Yes$ and $V15 \leq 3$
L58	$V1 = P$ and $V18 > 7.4$ and $V10 = No$ and $V13 = No$ and $V5 = III$ and $V12 = No$ and $V23 = Yes$ and $V15 \leq 3$
L59	$V8 = Yes$ and $V10 = Yes$ and $V13 = No$ and $V5 = III$ and $V12 = No$ and $V23 = Yes$ and $V15 \leq 3$
L60	$V18 \leq 8.32$ and $V8 = No$ and $V10 = Yes$ and $V13 = No$ and $V5 = III$ and $V12 = No$ and $V23 = Yes$ and $V15 \leq 3$
L61	$V13 = Yes$ and $V5 = III$ and $V12 = No$ and $V23 = Yes$ and $V15 \leq 3$
L62	$V5 = I - II$ and $V12 = No$ and $V23 = Yes$ and $V15 \leq 3$

Table 4.20: Low Risk Mortality Patterns $L63$-$L83$

Pattern	Pattern Description
L63	$V10 = Yes$ and $V5 = IV - VI$ and $V12 = No$ and $V23 = Yes$ and $V15 \leq 3$
L64	$V9 = No$ and $V17 \leq 0.049774$ and $V15 > 3$
L65	$V17 \leq 0.084034$ and $V10 = No$ and $V12 = Yes$
L66	$V2 \leq 58.1$ and $V17 > 0.084034$ and $V10 = No$ and $V12 = Yes$
L67	$V10 = Yes$ and $V12 = Yes$ and $V17 > 0.049774$ and $V15 > 3$ and $V16 \leq 0.007424$
L68	$V17 > 0.084034$ and $V11 = No$ and $V7 = No$ and $V16 \leq 0.004329$ and $V12 = No$
L69	$V23 = No$ and $V16 \leq 0.000373$ and $V7 = Yes$
L70	$V1 = B$ and $V5 = III$ and $V23 = No$ and $V18 \leq 15.55$ and $V10 = No$ and $V16 > 0.001736$
L71	$V2 \leq 67$ and $V1 = A$ and $V5 = III$ and $V23 = No$ and $V18 \leq 15.55$ and $V10 = No$ and $V16 > 0.001736$
L72	$V16 \leq 0.002959$ and $V5 = I - II$ and $V23 = No$ and $V18 \leq 15.55$ and $V10 = No$
L73	$V15 \leq 4$ and $V10 = Yes$ and $V16 > 0.001736$
L74	$V2 > 69.9$ and $V19 = Wednesday$ and $V15 > 4$ and $V10 = Yes$ and $V16 > 0.001736$
L75	$V19 = Friday$ and $V15 > 4$ and $V10 = Yes$ and $V16 > 0.001736$
L76	$V19 = Tuesday$ and $V15 > 4$ and $V10 = Yes$ and $V16 > 0.001736$
L77	$V2 \leq 74.9$ and $V18 \leq 12.75$ and $V12 = No$ and $V5 = III$ and $V10 = No$ and $V1 = C$ and $V22 = No$ and $V15 \leq 3$ and $V23 = No$ and $V16 > 0.007424$
L78	$V5 = I - II$ and $V10 = No$ and $V1 = C$ and $V22 = No$ and $V15 \leq 3$ and $V23 = No$ and $V16 > 0.007424$
L79	$V10 = Yes$ and $V1 = C$ and $V22 = No$ and $V15 \leq 3$ and $V23 = No$ and $V16 > 0.007424$
L80	$V10 = Yes$ and $V1 = T$ and $V22 = No$ and $V15 \leq 3$ and $V23 = No$ and $V16 > 0.007424$
L81	$V18$ $leq 9.95$ and $V10 = No$ and $V4 = Caucasian$ and $V1 = D$ and $V22 = No$ and $V15 \leq 3$ and $V23 = No$ and $V16 > 0.007424$
L82	$V10 = Yes$ and $V4 = Caucasian$ and $V1 = D$ and $V22 = No$ and $V15 \leq 3$ and $V23 = No$ and $V16 > 0.007424$
L83	$V4 = Other$ and $V1 = D$ and $V22 = No$ and $V15 \leq 3$ and $V23 = No$ and $V16 > 0.007424$

Table 4.21: High Risk Mortality Patterns $H1$-$H16$

Pattern	Pattern Description
H1	$V3 = Male$ and $V8 = Yes$ and $V17 > 0.031$ and $V16 \leq 0$ and $V23 = No$ and $V15 \leq 3$
H2	$V4 = AfricanAmerican$ and $V15 \leq 1$ and $V3 = Female$ and $V8 = Yes$ and $V17 \leq 0.031$ $V16 \leq 0$ and $V23 = No$
H3	$V15 > 1$ and $V3 = Female$ and $V8 = Yes$ and $V17 \leq 0.031$ $V16 \leq 0$ and $V23 = No$
H4	$V6 > 30.56$ and $V8 = No$ and $V17 > 0.031$ and $V16 \leq 0$ and $V23 = No$ and $V15 \leq 3$
H5	$V18 > 11.6$ and $V1 = A$ and $V16 > 0.0141$ and $V5 = III$ and $V9 = No$ and $V22 = No$ and $V12 = No$
H6	$V15 \leq 0$ and $V14 = Yes$
H7	$V2 > 74.2$ and $V8 = Yes$ and $V10 = No$ and $V15 > 1$ and $V1 = U$ and $V16 > 0.0141$ and $V5 = III$ and $V9 = No$ and $V22 = No$ and $V12 = No$
H8	$V8 = No$ and $V10 = No$ and $V15 > 1$ and $V1 = U$ and $V16 > 0.0141$ and $V5 = III$ and $V9 = No$ and $V22 = No$ and $V12 = No$
H9	$V18 > 14.2$ and $V13 = No$ and $V2 > 61.1$ and $V1 = G$ and $V16 > 0.0141$ and $V5 = III$ and $V9 = No$ and $V22 = No$ and $V12 = No$
H10	$V15 > 2$ and $V1 = N$ and $V16 > 0.0141$ and $V5 = III$ and $V9 = No$ and $V22 = No$ and $V12 = No$
H11	$V6 \leq 30.68$ and $V15 \leq 0$ and $V10 = No$
H12	$V15 > 0$ and $V10 = No$
H13	$V15 > 1$ and $V1 = Q$ and $V16 = 0.00141$ and $V5 = III$ and $V9 = No$ and $V22 = No$ and $V12 = No$
H14	$V1 = R$ and $V16 \leq 0.00141$ and $V5 = III$ and $V9 = No$ and $V22 = No$ and $V12 = No$
H15	$V1 = B$ and $V13 = No$ and $V10 = No$ and $V17 > 0.084034$ and $V5 = IV - VI$ and $V9 = No$ and $V22 = No$ and $V12 = No$ and $V16 > 0$ and $V23 = No$ and $V15 \leq 3$
H16	$V2 \leq 65.6$ and $V1 = M$ and $V13 = No$ and $V10 = No$ and $V17 > 0.084034$ and $V5 = IV - VI$ and $V9 = No$ and $V22 = No$ and $V12 = No$ and $V16 > 0$ and $V23 = No$ and $V15 \leq 3$

Table 4.22: High Risk Mortality Patterns $H17$-$H27$

Pattern	Pattern Description
H17	$V1 = A$ and $V13 = No$ and $V10 = No$ and $V17 > 0.084034$ and $V5 = IV - VI$ and $V9 = No$ and $V22 = No$ and $V12 = No$ and $V16 > 0$ and $V23 = No$ and $V15 \leq 3$
H18	$V6 \leq 28.9$ and $V1 = J$ and $V13 = No$ and $V10 = No$ and $V17 > 0.084034$ and $V5 = IV - VI$ and $V9 = No$ and $V22 = No$ and $V12 = No$ and $V16 > 0$ and $V23 = No$ and $V15 \leq 3$
H19	$V1 = L$ and $V13 = No$ and $V10 = No$ and $V17 > 0.084034$ and $V5 = IV - VI$ and $V9 = No$ and $V22 = No$ and $V12 = No$ and $V16 > 0$ and $V23 = No$ and $V15 \leq 3$
H20	$V1 = H$ and $V13 = No$ and $V10 = No$ and $V17 > 0.084034$ and $V5 = IV - VI$ and $V9 = No$ and $V22 = No$ and $V12 = No$ and $V16 > 0$ and $V23 = No$ and $V15 \leq 3$
H21	$V14 = Yes$ and $V8 = Yes$ and $V15 \leq 1$ and $V1 = U$ and $V13 = No$ and $V10 = No$ and $V17 > 0.084034$ and $V5 = IV - VI$ and $V9 = No$ and $V22 = No$ and $V12 = No$ and $V16 > 0$ and $V23 = No$
H22	$V8 = No$ and $V15 \leq 1$ and $V1 = U$ and $V13 = No$ and $V10 = No$ and $V17 > 0.084034$ and $V5 = IV - VI$ and $V9 = No$ and $V22 = No$ and $V12 = No$ and $V16 > 0$ and $V23 = No$
H23	$V15 > 1$ and $V1 = U$ and $V13 = No$ and $V10 = No$ and $V17 > 0.084034$ and $V5 = IV - VI$ and $V9 = No$ and $V22 = No$ and $V12 = No$ and $V16 > 0$ and $V23 = No$
H24	$V1 = G$ and $V13 = No$ and $V10 = No$ and $V17 > 0.084034$ and $V5 = IV - VI$ and $V9 = No$ and $V22 = No$ and $V12 = No$ and $V16 > 0$ and $V23 = No$ and $V15 \leq 3$
H25	$V1 = O$ and $V13 = No$ and $V10 = No$ and $V17 > 0.084034$ and $V5 = IV - VI$ and $V9 = No$ and $V22 = No$ and $V12 = No$ and $V16 > 0$ and $V23 = No$ and $V15 \leq 3$
H26	$V1 = K$ and $V13 = No$ and $V10 = No$ and $V17 > 0.084034$ and $V5 = IV - VI$ and $V9 = No$ and $V22 = No$ and $V12 = No$ and $V16 > 0$ and $V23 = No$ and $V15 \leq 3$
H27	$V1 = W$ and $V13 = No$ and $V10 = No$ and $V17 > 0.084034$ and $V5 = IV - VI$ and $V9 = No$ and $V22 = No$ and $V12 = No$ and $V16 > 0$ and $V23 = No$ and $V15 \leq 3$

Table 4.23: High Risk Mortality Patterns $H28$-$H37$

Pattern	Pattern Description
H28	$V1 = C$ and $V13 = No$ and $V10 = No$ and $V17 > 0.084034$ and $V5 = IV - VI$ and $V9 = No$ and $V22 = No$ and $V12 = No$ and $V16 > 0$ and $V23 = No$ and $V15 \leq 3$
H29	$V1 = N$ and $V13 = No$ and $V10 = No$ and $V17 > 0.084034$ and $V5 = IV - VI$ and $V9 = No$ and $V22 = No$ and $V12 = No$ and $V16 > 0$ and $V23 = No$ and $V15 \leq 3$
H30	$V1 = V$ and $V13 = No$ and $V10 = No$ and $V17 > 0.084034$ and $V5 = IV - VI$ and $V9 = No$ and $V22 = No$ and $V12 = No$ and $V16 > 0$ and $V23 = No$ and $V15 \leq 3$
H31	$V1 = Q$ and $V13 = No$ and $V10 = No$ and $V17 > 0.084034$ and $V5 = IV - VI$ and $V9 = No$ and $V22 = No$ and $V12 = No$ and $V16 > 0$ and $V23 = No$ and $V15 \leq 3$
H32	$V1 = E$ and $V13 = No$ and $V10 = No$ and $V17 > 0.084034$ and $V5 = IV - VI$ and $V9 = No$ and $V22 = No$ and $V12 = No$ and $V16 > 0$ and $V23 = No$ and $V15 \leq 3$
H33	$V1 = P$ and $V13 = No$ and $V10 = No$ and $V17 > 0.084034$ and $V5 = IV - VI$ and $V9 = No$ and $V22 = No$ and $V12 = No$ and $V16 > 0$ and $V23 = No$ and $V15 \leq 3$
H34	$V1 = T$ and $V13 = No$ and $V10 = No$ and $V17 > 0.084034$ and $V5 = IV - VI$ and $V9 = No$ and $V22 = No$ and $V12 = No$ and $V16 > 0$ and $V23 = No$ and $V15 \leq 3$
H35	$V1 = S$ and $V13 = No$ and $V10 = No$ and $V17 > 0.084034$ and $V5 = IV - VI$ and $V9 = No$ and $V22 = No$ and $V12 = No$ and $V16 > 0$ and $V23 = No$ and $V15 \leq 3$
H36	$V1 = I$ and $V13 = No$ and $V10 = No$ and $V17 > 0.084034$ and $V5 = IV - VI$ and $V9 = No$ and $V22 = No$ and $V12 = No$ and $V16 > 0$ and $V23 = No$ and $V15 \leq 3$
H37	$V1 = D$ and $V13 = No$ and $V10 = No$ and $V17 > 0.084034$ and $V5 = IV - VI$ and $V9 = No$ and $V22 = No$ and $V12 = No$ and $V16 > 0$ and $V23 = No$ and $V15 \leq 3$

Table 4.24: High Risk Mortality Patterns $H38$-$H50$

Pattern	Pattern Description
H38	$V1 = F$ and $V13 = No$ and $V10 = No$ and $V17 > 0.084034$ and $V5 = IV - VI$ and $V9 = No$ and $V22 = No$ and $V12 = No$ and $V16 > 0$ and $V23 = No$ and $V15 \leq 3$
H39	$V1 = R$ and $V13 = No$ and $V10 = No$ and $V17 > 0.084034$ and $V5 = IV - VI$ and $V9 = No$ and $V22 = No$ and $V12 = No$ and $V16 > 0$ and $V23 = No$ and $V15 \leq 3$
H40	$V2 \leq 73.9$ and $V15 > 2$ and $V10 = Yes$ and $V17 > 0.084034$ and $V5 = IV - VI$ and $V9 = No$ and $V22 = No$ and $V12 = No$ and $V16 > 0$ and $V23 = No$
H41	$V13 = No$ and $V9 = Yes$ and $V22 = No$ and $V12 = No$ and $V16 > 0$ and $V23 = No$ and $V15 \leq 3$
H42	$V22 = Yes$ and $V12 = No$ and $V16 > 0$ and $V23 = No$ and $V15 \leq 3$
H43	$V2 > 77.4$ and $V1 = B$ and $V18 > 7.4$ and $V10 = No$ and $V13 = No$ and $V5 = III$ and $V12 = No$ and $V23 = Yes$ and $V15 \leq 3$
H44	$V15 > 1$ and $V1 = M$ $V18 > 7.4$ and $V10 = No$ and $V13 = No$ and $V5 = III$ and $V12 = No$ and $V23 = Yes$
H45	$V1 = A$ and $V18 > 7.4$ and $V10 = No$ and $V13 = No$ and $V5 = III$ and $V12 = No$ and $V23 = Yes$ and $V15 \leq 3$
H46	$V18 \leq 9.75$ and $V15 > 0$ and $V4 = Caucasian$ and $V11 = No$ and $V1 = J$ and $V18 > 7.4$ and $V10 = No$ and $V13 = No$ and $V5 = III$ and $V12 = No$ and $V23 = Yes$
H47	$V4 = African - American$ and $V11 = No$ and $V1 = J$ and $V18 > 7.4$ and $V10 = No$ and $V13 = No$ and $V5 = III$ and $V12 = No$ and $V23 = Yes$ and $V15 \leq 3$
H48	$V4 = Other$ and $V11 = No$ and $V1 = J$ and $V18 > 7.4$ AND $V10 = No$ and $V13 = No$ and $V5 = III$ and $V12 = No$ and $V23 = Yes$ and $V15 \leq 3$
H49	$V7 = No$ and $V1 = H$ and $V18 > 7.4$ and $V10 = No$ and $V13 = No$ and $V5 = III$ and $V12 = No$ and $V23 = Yes$ and $V15 \leq 3$
H50	$V11 = Yes$ and $V7 = Yes$ and $V1 = H$ and $V18 > 7.4$ and $V10 = No$ and $V13 = No$ and $V5 = III$ and $V12 = No$ and $V23 = Yes$ and $V15 \leq 3$

Table 4.25: High Risk Mortality Patterns $H51$-$H70$

Pattern	Pattern Description
H51	$V4 = AfricanAmerican$ and $V11 = Yes$ and $V18 > 7.55$ and $V1 = U$
H52	$V4 = Caucasian$ and $V11 = No$ and $V18 > 7.55$ and $V1 = U$
H53	$V4 = Other$ and $V11 = No$ and $V18 > 7.55$ and $V1 = U$
H54	$V2 > 64.4$ and $V1 = G$ and $V18 > 7.4$ and $V10 = No$ and $V13 = No$ and $V5 = III$ and $V12 = No$ and $V23 = Yes$ and $V15 \leq 3$
H55	$V1 = K$ and $V18 > 7.4$ and $V10 = No$ and $V13 = No$ and $V5 = III$ and $V12 = No$ and $V23 = Yes$ and $V15 \leq 3$
H56	$V1 = W$ and $V18 > 7.4$ and $V10 = No$ and $V13 = No$ and $V5 = III$ and $V12 = No$ and $V23 = Yes$ and $V15 \leq 3$
H57	$V1 = C$ and $V18 > 7.4$ and $V10 = No$ and $V13 = No$ and $V5 = III$ and $V12 = No$ and $V23 = Yes$ and $V15 \leq 3$
H58	$V1 = N$ and $V18 > 7.4$ and $V10 = No$ and $V13 = No$ and $V5 = III$ and $V12 = No$ and $V23 = Yes$ and $V15 \leq 3$
H59	$V1 = V$ and $V18 > 7.4$ and $V10 = No$ and $V13 = No$ and $V5 = III$ and $V12 = No$ and $V23 = Yes$ and $V15 \leq 3$
H60	$V1 = Q$ and $V18 > 7.4$ and $V10 = No$ and $V13 = No$ and $V5 = III$ and $V12 = No$ and $V23 = Yes$ and $V15 \leq 3$
H61	$V1 = E$ and $V18 > 7.4$ and $V10 = No$ and $V13 = No$ and $V5 = III$ and $V12 = No$ and $V23 = Yes$ and $V15 \leq 3$
H62	$V1 = T$ and $V18 > 7.4$ and $V10 = No$ and $V13 = No$ and $V5 = III$ and $V12 = No$ and $V23 = Yes$ and $V15 \leq 3$
H63	$V1 = S$ and $V18 > 7.4$ and $V10 = No$ and $V13 = No$ and $V5 = III$ and $V12 = No$ and $V23 = Yes$ and $V15 \leq 3$
H64	$V1 = I$ and $V18 > 7.4$ and $V10 = No$ and $V13 = No$ and $V5 = III$ and $V12 = No$ and $V23 = Yes$ and $V15 \leq 3$
H65	$V1 = D$ and $V18 > 7.4$ and $V10 = No$ and $V13 = No$ and $V5 = III$ and $V12 = No$ and $V23 = Yes$ and $V15 \leq 3$
H66	$V1 = F$ and $V18 > 7.4$ and $V10 = No$ and $V13 = No$ and $V5 = III$ and $V12 = No$ and $V23 = Yes$ and $V15 \leq 3$
H67	$V1 = R$ and $V18 > 7.4$ and $V10 = No$ and $V13 = No$ and $V5 = III$ and $V12 = No$ and $V23 = Yes$ and $V15 \leq 3$
H68	$V18 > 8.32$ and $V8 = No$ and $V10 = Yes$ and $V13 = No$ and $V5 = III$ and $V12 = No$ and $V23 = Yes$ and $V15 \leq 3$
H69	$V10 = No$ and $V5 = IV - VI$ and $V12 = No$ and $V23 = Yes$ and $V15 \leq 3$
H70	$V9 = Yes$ and $V17 \leq 0.049774$ and $V15 > 3$

Table 4.26: High Risk Mortality Patterns $H71$-$H85$

Pattern	Pattern Description
H71	$V2 > 58.1$ and $V17 > 0.084034$ and $V10 = No$ and $V12 = Yes$
H72	$V11 = Yes$ and $V7 = No$ and $V17 > 0.049774$ and $V15 > 3$ and $V16 \leq 0.007424$
H73	$V17 \leq 0.084034$ and $V11 = No$ and $V7 = No$ and $V16 \leq 0.004329$ and $V12 = No$
H74	$V23 = Yes$ and $V16 \leq 0.000373$ and $V7 = Yes$
H75	$V16 \leq 0.001736$
H76	$V1 = M$ and $V5 = III$ and $V23 = No$ and $V18 \leq 15.55$ and $V10 = No$ and $V16 > 0.001736$
H77	$V2 > 67$ and $V1 = A$ and $V5 = III$ and $V23 = No$ and $V18 \leq 15.55$ and $V10 = No$ and $V16 > 0.001736$
H78	$V1 = J$ and $V5 = III$ and $V23 = No$ and $V18 \leq 15.55$ and $V10 = No$ and $V16 > 0.001736$
H79	$V1 = L$ and $V5 = III$ and $V23 = No$ and $V18 \leq 15.55$ and $V10 = No$ and $V16 > 0.001736$
H80	$V1 = H$ and $V5 = III$ and $V23 = No$ and $V18 \leq 15.55$ and $V10 = No$ and $V16 > 0.001736$
H81	$V1 = U$ and $V5 = III$ and $V23 = No$ and $V18 \leq 15.55$ and $V10 = No$ and $V16 > 0.001736$
H82	$V1 = G$ and $V5 = III$ and $V23 = No$ and $V18 \leq 15.55$ and $V10 = No$ and $V16 > 0.001736$
H83	$V1 = O$ and $V5 = III$ and $V23 = No$ and $V18 \leq 15.55$ and $V10 = No$ and $V16 > 0.001736$
H84	$V1 = K$ and $V5 = III$ and $V23 = No$ and $V18 \leq 15.55$ and $V10 = No$ and $V16 > 0.001736$
H85	$V1 = W$ and $V5 = III$ and $V23 = No$ and $V18 \leq 15.55$ and $V10 = No$ and $V16 > 0.001736$

Table 4.27: High Risk Mortality Patterns $H86$-$H104$

Pattern	Pattern Description
H86	$V1 = C$ and $V5 = III$ and $V23 = No$ and $V18 \leq 15.55$ and $V10 = No$ and $V16 > 0.001736$
H87	$V1 = N$ and $V5 = III$ and $V23 = No$ and $V18 \leq 15.55$ and $V10 = No$ and $V16 > 0.001736$
H88	$V1 = V$ and $V5 = III$ and $V23 = No$ and $V18 \leq 15.55$ and $V10 = No$ and $V16 > 0.001736$
H89	$V1 = Q$ and $V5 = III$ and $V23 = No$ and $V18 \leq 15.55$ and $V10 = No$ and $V16 > 0.001736$
H90	$V1 = E$ and $V5 = III$ and $V23 = No$ and $V18 \leq 15.55$ and $V10 = No$ and $V16 > 0.001736$
H91	$V1 = P$ and $V5 = III$ and $V23 = No$ and $V18 \leq 15.55$ and $V10 = No$ and $V16 > 0.001736$
H92	$V1 = T$ and $V5 = III$ and $V23 = No$ and $V18 \leq 15.55$ and $V10 = No$ and $V16 > 0.001736$
H93	$V1 = S$ and $V5 = III$ and $V23 = No$ and $V18 \leq 15.55$ and $V10 = No$ and $V16 > 0.001736$
H94	$V1 = I$ and $V5 = III$ and $V23 = No$ and $V18 \leq 15.55$ and $V10 = No$ and $V16 > 0.001736$
H95	$V1 = D$ and $V5 = III$ and $V23 = No$ and $V18 \leq 15.55$ and $V10 = No$ and $V16 > 0.001736$
H96	$V1 = F$ and $V5 = III$ and $V23 = No$ and $V18 \leq 15.55$ and $V10 = No$ and $V16 > 0.001736$
H97	$V1 = R$ and $V5 = III$ and $V23 = No$ and $V18 \leq 15.55$ and $V10 = No$ and $V16 > 0.001736$
H98	$V16 > 0.002959$ and $V5 = I - II$ and $V23 = No$ and $V18 \leq 15.55$ and $V10 = No$
H99	$V5 = IV - VI$ and $V23 = No$ and $V18 \leq 15.55$ and $V10 = No$ and $V16 > 0.001736$
H100	$V23 = Yes$ and $V18 \leq 15.55$ and $V10 = No$ and $V16 > 0.001736$
H101	$V18 > 15.55$ and $V10 = No$ and $V16 > 0.001736$
H102	$V19 = Thursday$ and $V15 > 4$ and $V10 = Yes$ and $V16 > 0.001736$
H103	$V2 \leq 69.9$ and $V19 = Wednesday$ and $V15 > 4$ and $V10 = Yes$ and $V16 > 0.001736$
H104	$V19 = Monday$ and $V15 > 4$ and $V10 = Yes$ and $V16 > 0.001736$

Table 4.28: High Risk Mortality Patterns $H105$-$H120$

Pattern	Pattern Description
H105	$V16 > 0.004329$ and $V12 = No$ and $V17 > 0.049774$ and $V15 > 3$
H106	$V1 = B$ and $V22 = No$ and $V15 \leq 3$ and $V23 = No$ and $V16 > 0.007424$
H107	$V1 = M$ and $V22 = No$ and $V15 \leq 3$ and $V23 = No$ and $V16 > 0.007424$
H108	$V1 = A$ and $V22 = No$ and $V15 \leq 3$ and $V23 = No$ and $V16 > 0.007424$
H109	$V1 = J$ and $V22 = No$ and $V15 \leq 3$ and $V23 = No$ and $V16 > 0.007424$
H110	$V1 = L$ and $V22 = No$ and $V15 \leq 3$ and $V23 = No$ and $V16 > 0.007424$
H111	$V1 = H$ and $V22 = No$ and $V15 \leq 3$ and $V23 = No$ and $V16 > 0.007424$
H112	$V1 = U$ and $V22 = No$ and $V15 \leq 3$ and $V23 = No$ and $V16 > 0.007424$
H113	$V1 = G$ and $V22 = No$ and $V15 \leq 3$ and $V23 = No$ and $V16 > 0.007424$
H114	$V1 = O$ and $V22 = No$ and $V15 \leq 3$ and $V23 = No$ and $V16 > 0.007424$
H115	$V1 = K$ and $V22 = No$ and $V15 \leq 3$ and $V23 = No$ and $V16 > 0.007424$
H116	$V1 = W$ and $V22 = No$ and $V15 \leq 3$ and $V23 = No$ and $V16 > 0.007424$
H117	$V12 = Yes$ and $V5 = III$ and $V10 = No$ and $V1 = C$ and $V22 = No$ and $V15 \leq 3$ and $V23 = No$ and $V16 > 0.007424$
H118	$V2 > 74.9$ and $V18 \leq 12.75$ and $V12 = No$ and $V5 = III$ and $V10 = No$ and $V1 = C$ and $V22 = No$ and $V15 \leq 3$ and $V23 = No$ and $V16 > 0.007424$
H119	$V18 > 12.75$ and $V12 = No$ and $V5 = III$ and $V10 = No$ and $V1 = C$ and $V22 = No$ and $V15 \leq 3$ and $V23 = No$ and $V16 > 0.007424$
H120	$V5 = IV - VI$ and $V10 = No$ and $V1 = C$ and $V22 = No$ and $V15 \leq 3$ and $V23 = No$ and $V16 > 0.007424$

Table 4.29: High Risk Mortality Patterns $H121$-$H135$

Pattern	Pattern Description
H121	$V1 = N$ and $V22 = No$ and $V15 \leq 3$ and $V23 = No$ and $V16 > 0.007424$
H122	$V1 = V$ and $V22 = No$ and $V15 \leq 3$ and $V23 = No$ and $V16 > 0.007424$
H123	$V1 = Q$ and $V22 = No$ and $V15 \leq 3$ and $V23 = No$ and $V16 > 0.007424$
H124	$V1 = E$ and $V22 = No$ and $V15 \leq 3$ and $V23 = No$ and $V16 > 0.007424$
H125	$V1 = P$ and $V22 = No$ and $V15 \leq 3$ and $V23 = No$ and $V16 > 0.007424$
H126	$V10 = No$ and $V1 = T$ and $V22 = No$ and $V15 \leq 3$ and $V23 = No$ and $V16 > 0.007424$
H127	$V1 = S$ and $V22 = No$ and $V15 \leq 3$ and $V23 = No$ and $V16 > 0.007424$
H128	$V1 = I$ and $V22 = No$ and $V15 \leq 3$ and $V23 = No$ and $V16 > 0.007424$
H129	$V18 > 9.95$ and $V10 = No$ and $V4 = Caucasian$ and $V1 = D$ and $V22 = No$ and $V15 \leq 3$ and $V23 = No$ and $V16 > 0.007424$
H130	$V4 = African - American$ and $V1 = D$ and $V22 = No$ and $V15 \leq 3$ and $V23 = No$ and $V16 > 0.007424$
H131	$V1 = F$ and $V22 = No$ and $V15 \leq 3$ and $V23 = No$ and $V16 > 0.007424$
H132	$V1 = R$ and $V22 = No$ and $V15 \leq 3$ and $V23 = No$ and $V16 > 0.007424$
H133	$V22 = Yes$ and $V15 \leq 3$ and $V23 = No$ and $V16 > 0.007424$
H134	$V15 > 3$ and $V23 = No$ and $V16 > 0.007424$
H135	$V23 = Yes$ and $V16 > 0.007424$

4.2 Prediction of Low, Medium, and High Risk Mortality

4.2.1 Identification of Low, Medium, and High Risk Patients

As discussed in Section 4.1, we are able to accurately predict low and high risk mortality of patients who had a surgery in Cleveland Clinic between January 2005 and September 2010. In this section we extend our investigation to predict low, medium, and high risk mortality in Surgery Timing dataset based on the *Mortality RSI* values ranging from -4.4 to 4.86. The initial step of our investigation is to discretize the *Mortality RSI* values as follows:

- A patient is labeled as "low risk" patient if the patient's corresponding *Mortality RSI* value is between -4.4 to -1, i.e., the probability that the patient's dies within 30 days after the surgery is between 1.21% and 26%. Note that these are the same low risk patients included in Surgery Timing LH dataset.

- A patient is labeled as "medium risk" patient if the patient's corresponding *Mortality RSI* value is between -0.5 to 0.5, i.e., the probability that the patient's dies within 30 days after the surgery is between 37.75% and 62.2%.

- A patient is labeled as "high risk" patient if the patient's corresponding *Mortality RSI* value is between 1 and 4.86, i.e., the probability that the patient's dies within 30 days after the surgery is between 73.1% and 99.2%. Note that these are the same high risk patients included in Surgery Timing LH dataset.

Due to high misclassification rate, we removed the patients whose *Mortality RSI* values are in interval (-1,-0.5) or in interval (0.5,1). The resulting dataset, referred to as "Surgery Timing LMH", contains 9,559 low risk, 15,217 medium risk, and 1,469 high risk patients and their corresponding features in Table 3.1.

Table 4.30 and Figure 4.23 show the distribution of the low risk, medium risk, and high risk patients based on their surgical procedure. Table 4.31 and Figures 4.24-4.28 give the distribution of the low risk, medium risk, and high risk patients based on their age, gender, race, ASA physical status, and BMI, respectively. Distribution of underlying health conditions among low risk, medium risk, and high risk patients are presented in Table 4.32 and Figures 4.29-4.35. In Table 4.33 and Figures 4.36-4.38, we give the distribution of baseline Charlson index, ccsMort30rate, and ccsComplication rate among low risk, medium risk, and high risk patients, respectively. Table 4.34 and Figures 4.39-4.44 show the distribution of hour, day of week, month, moon phase, 30-day mortality and in-hospital complication of low risk, medium risk, and high risk patients, respectively.

Table 4.30: Low Risk, Medium Risk, and High Risk Patients Data Characteristics - Surgery Type

Class	**# of Patients**
Low Risk	9,559
Medium Risk	15,217
High Risk	1,469
Surgery Type	**# of Patients**
Surgery A	780
Sugery B	3,265
Surgery C	1,835
Surgery D	281
Surgery E	296
Surgery F	401
Surgery G	2,034
Surgery H	2,162
Surgery I	698
Surgery J	1,999
Surgery K	468
Surgery L	412
Surgery M	2,093
Surgery N	497
Surgery O	1,820
Surgery P	840
Surgery Q	1,042
Surgery R	430
Surgery S	420
Surgery T	515
Surgery U	2,186
Surgery V	1,289
Surgery W	482

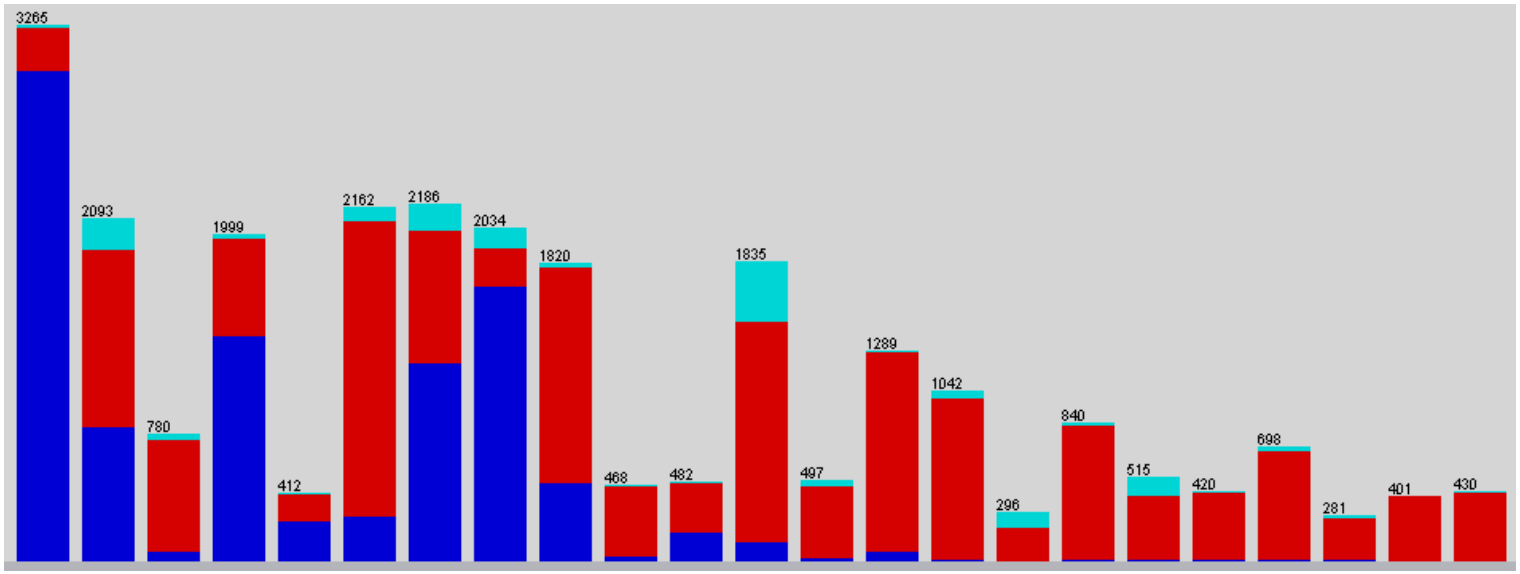

Figure 4.23: Surgery Type Distribution among Low Risk, Medium Risk, and High Risk Patients

Table 4.31: Low Risk, Medium Risk, and High Risk Patients Data Characteristics - Age, Gender, Race, ASA Physical Status, BMI

Age	Min Val. = 1, Max Val. = 90, Mean = 56.889, St. Dev. = 15.212, Missing Val. = 2, Distinct Val. = 807, Unique Val. = 52
Gender	
Male	14,604
Female	11,639
Race	
Caucasian	21,676
African-American	3,163
Other	1,013
ASA Physical Status	
I-II	14,674
III	10,739
IV-VI	826
BMI	Min Val. = 2.15, Max Val. = 92.59, Mean = 29.463, St. Dev. = 7.289, Missing Val = 2,732, Distinct Val. = 3,271, Unique Val. = 735

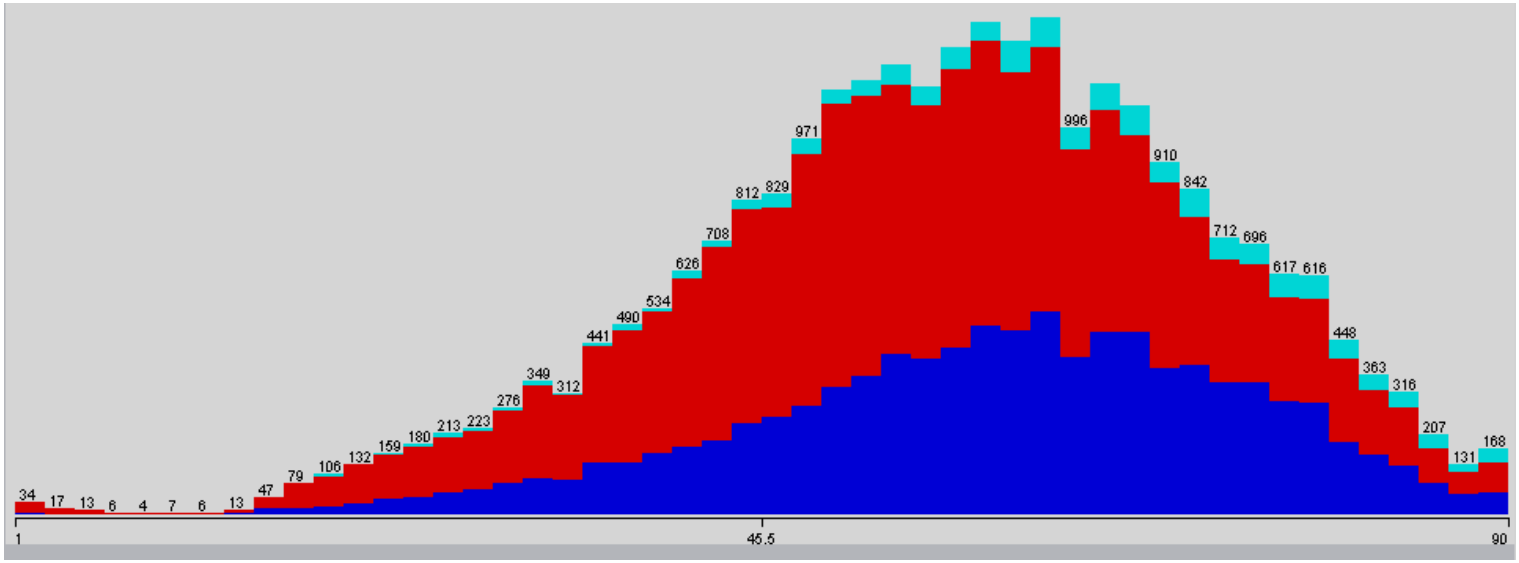

Figure 4.24: Age Distribution among Low Risk, Medium Risk, and High Risk Patients

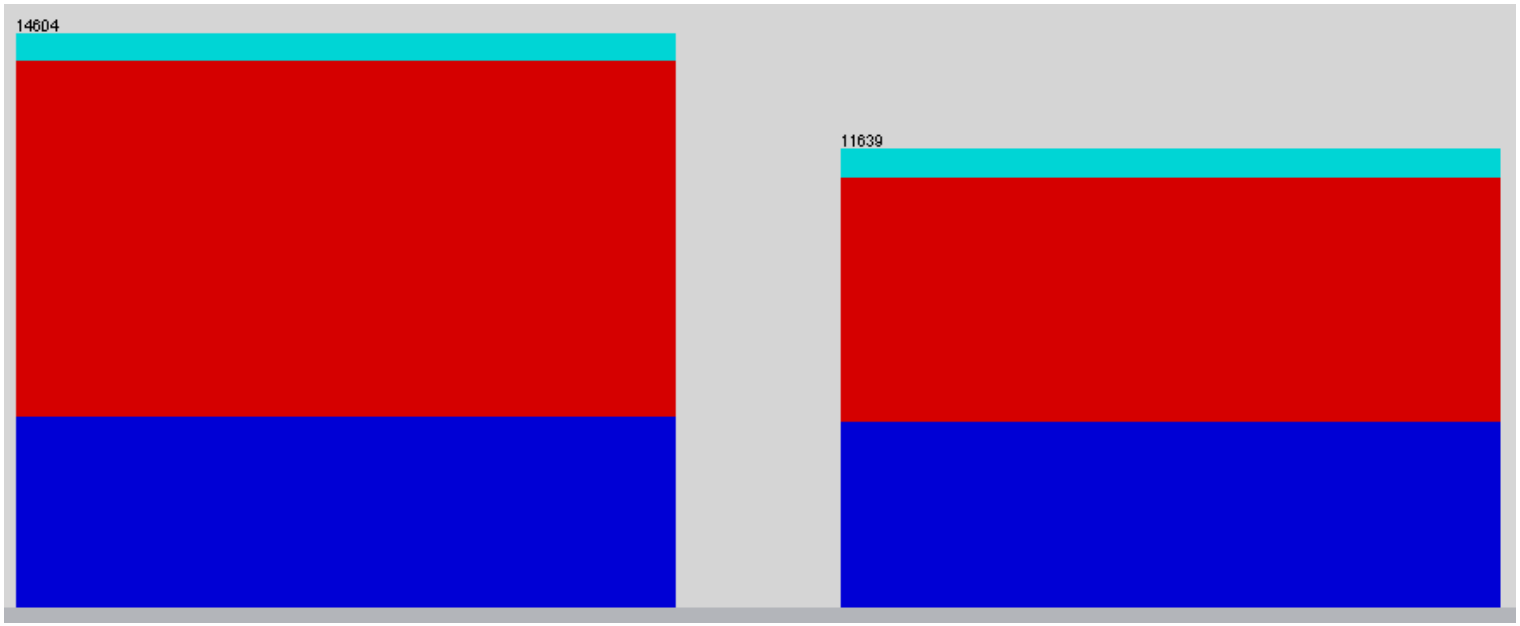

Figure 4.25: Gender Distribution among Low Risk, Medium Risk, and High Risk Patients

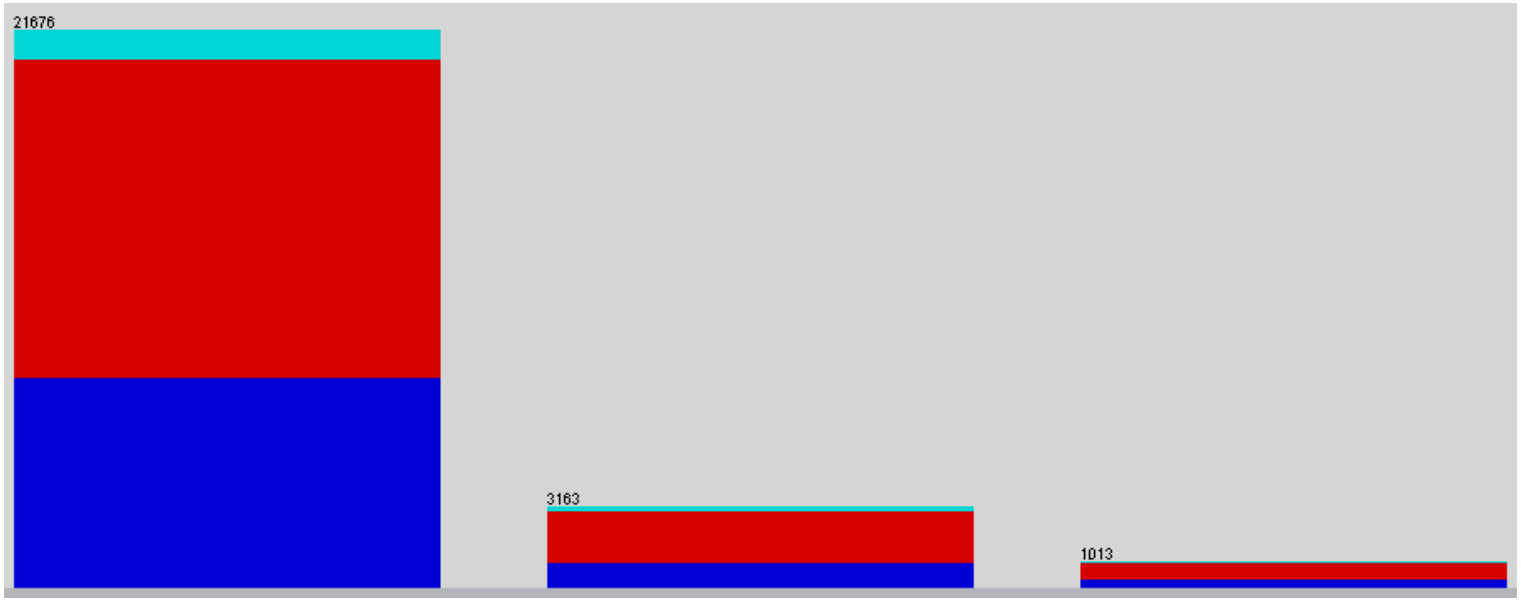

Figure 4.26: Race Distribution among Low Risk, Medium Risk, and High Risk Patients

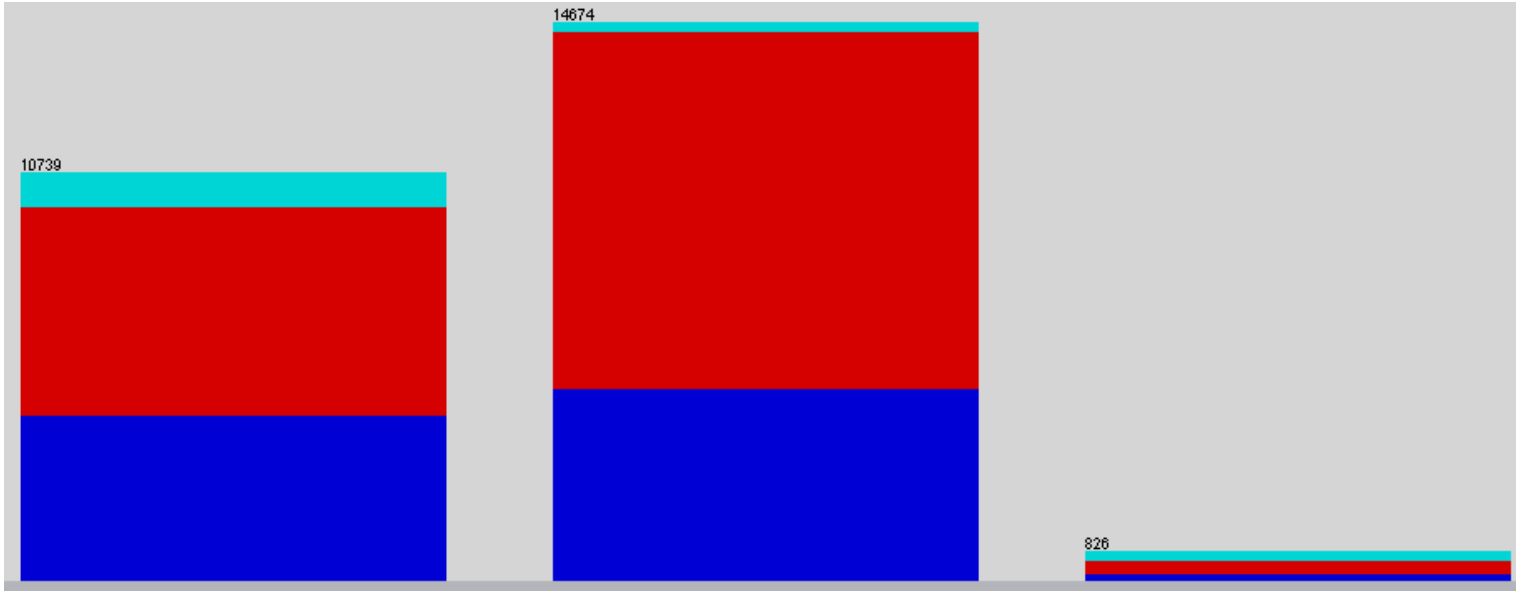

Figure 4.27: ASA Physical Status Distribution among Low Risk, Medium Risk, and High Risk Patients

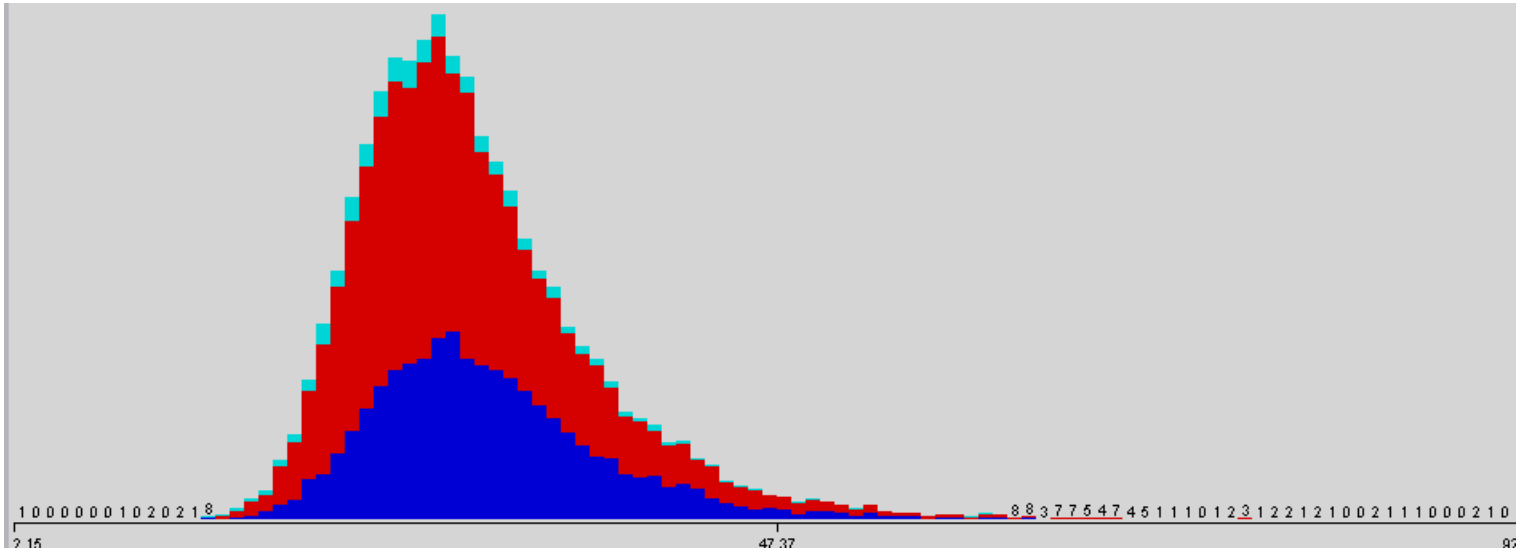

Figure 4.28: BMI Distribution among Low Risk, Medium Risk, and High Risk Patients

Table 4.32: Low Risk, Medium Risk, and High Risk Patients Data Characteristics - Underlying Health Conditions

Baseline Cancer	
No	18,068
Yes	8,177
Baseline CVD	
Yes	12,660
No	13,585
Baseline Dementia	
No	26,055
Yes	190
Baseline Diabetes	
No	23,011
Yes	3,234
Baseline Digestive	
Yes	5,660
No	20,585
Baseline Osteoart	
Yes	5,264
No	20,981
Baseline Psych	
No	23,859
Yes	2,386
Baseline Pulmonary	
No	23,418
Yes	2,827

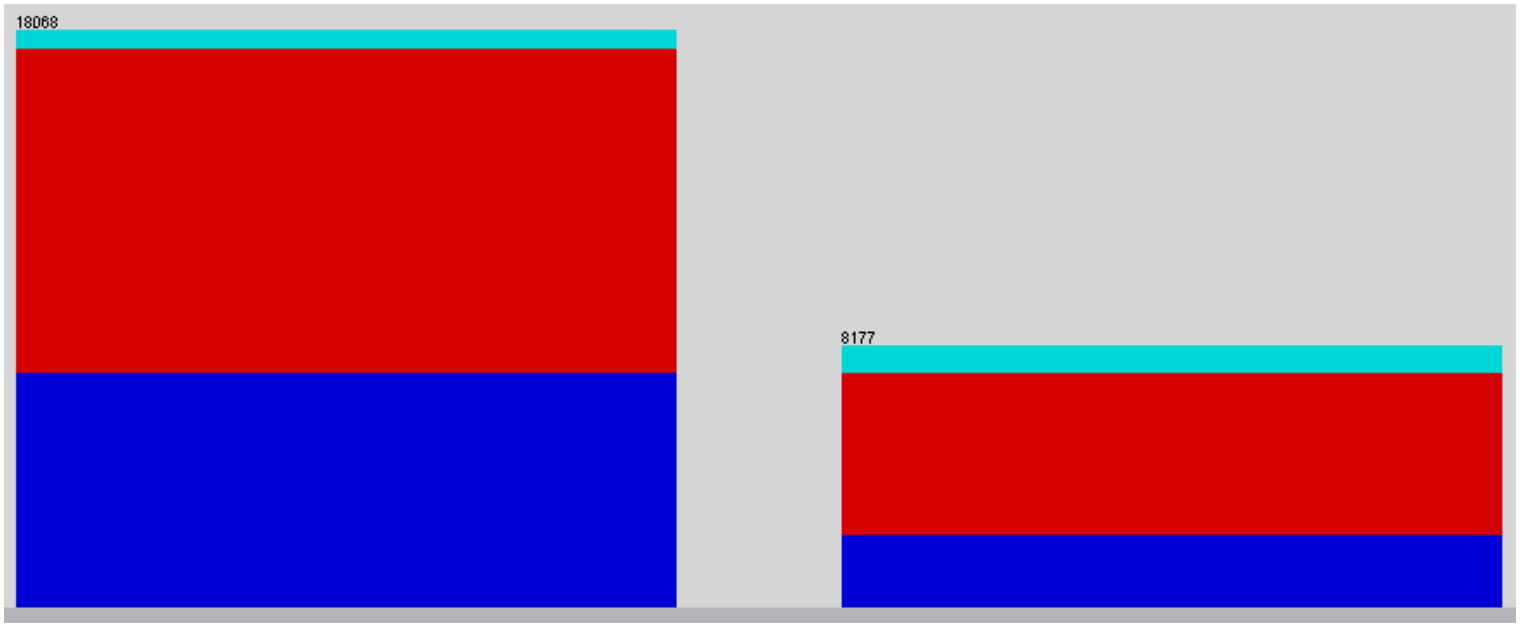

Figure 4.29: Baseline Cancer Distribution among Low Risk, Medium Risk, and High Risk Patients

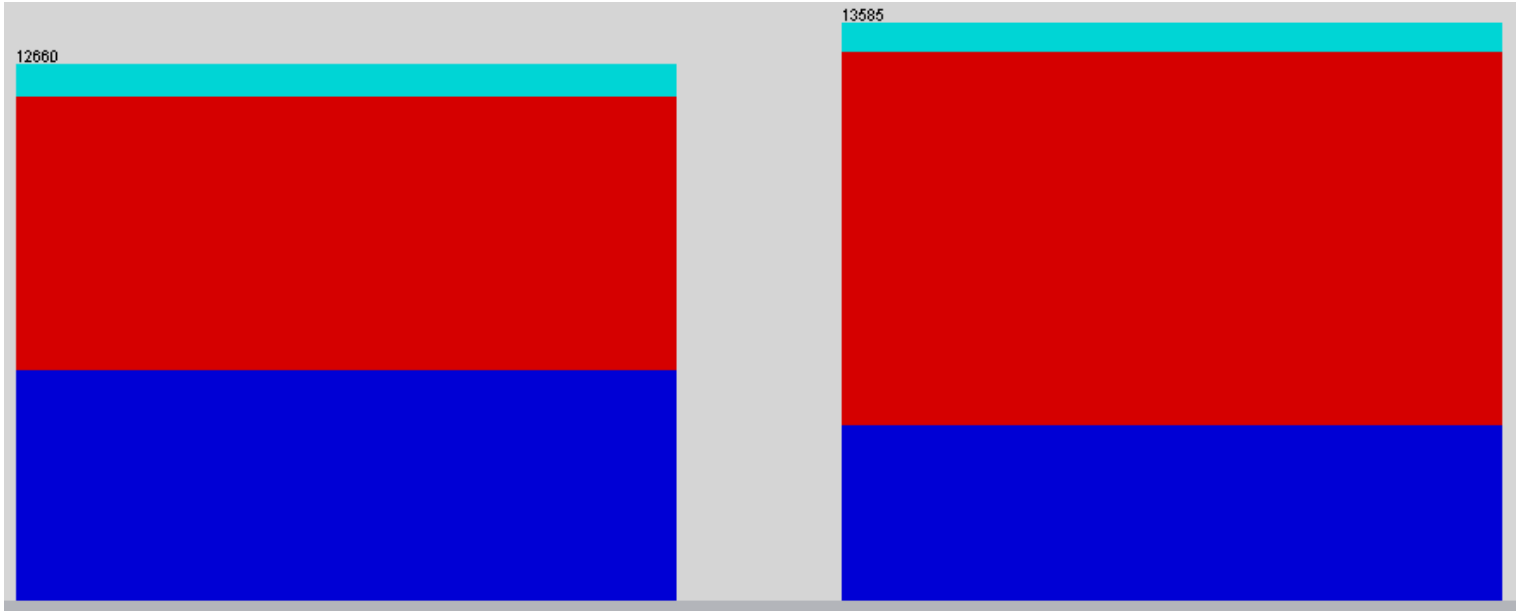

Figure 4.30: Baseline CVD Distribution among Low Risk, Medium Risk, and High Risk Patients

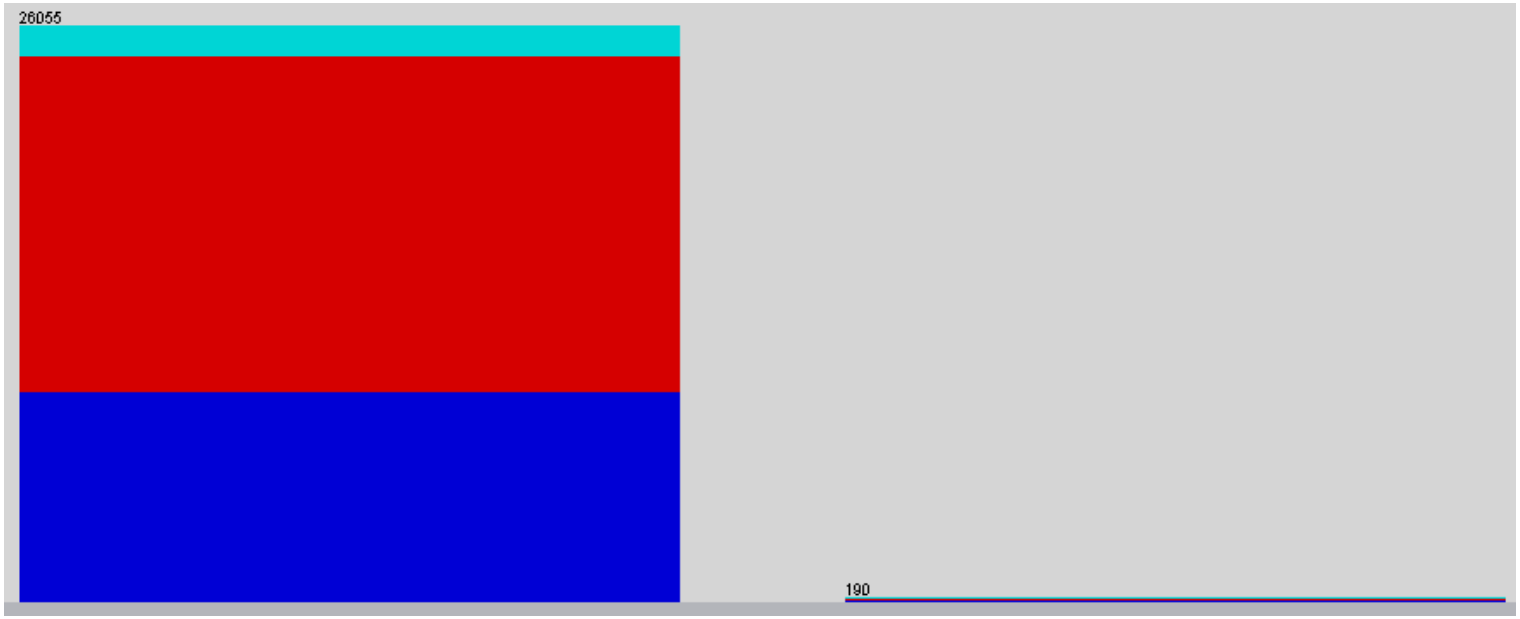

Figure 4.31: Baseline Dementia Distribution among Low Risk, Medium Risk, and High Risk Patients

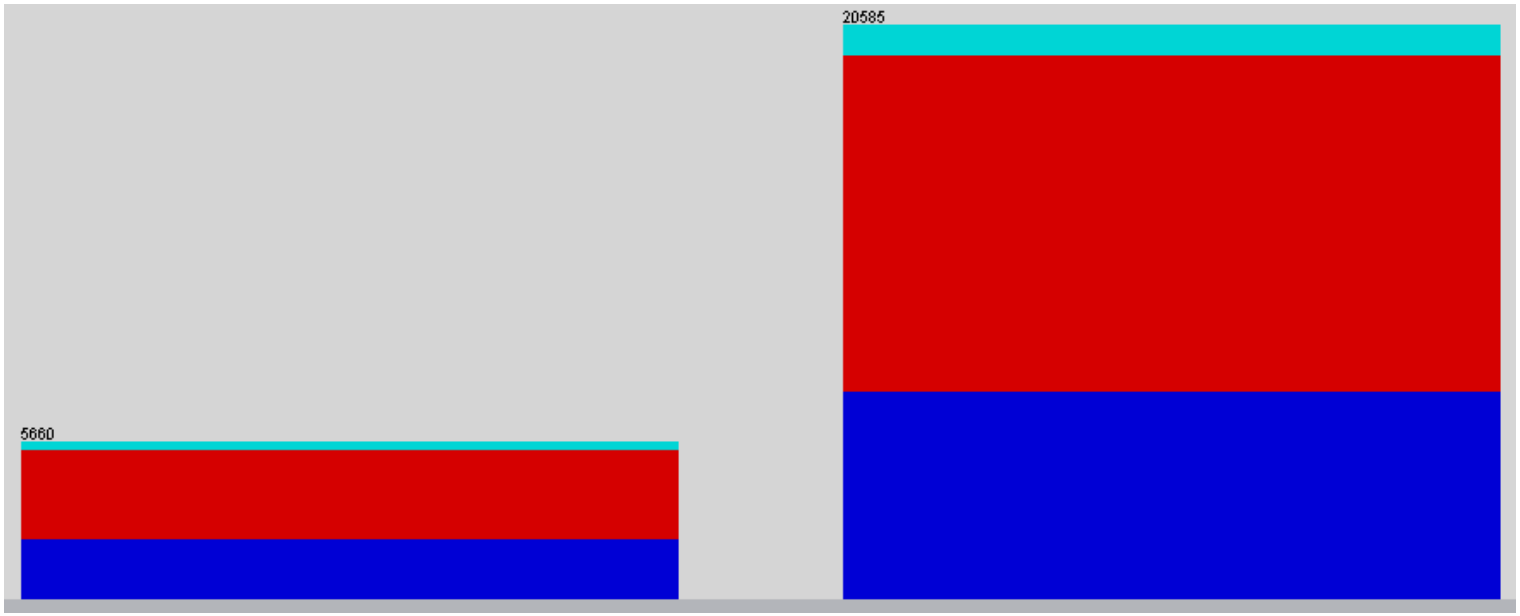

Figure 4.32: Baseline Digestive Disorder Distribution among Low Risk, Medium Risk, and High Risk Patients

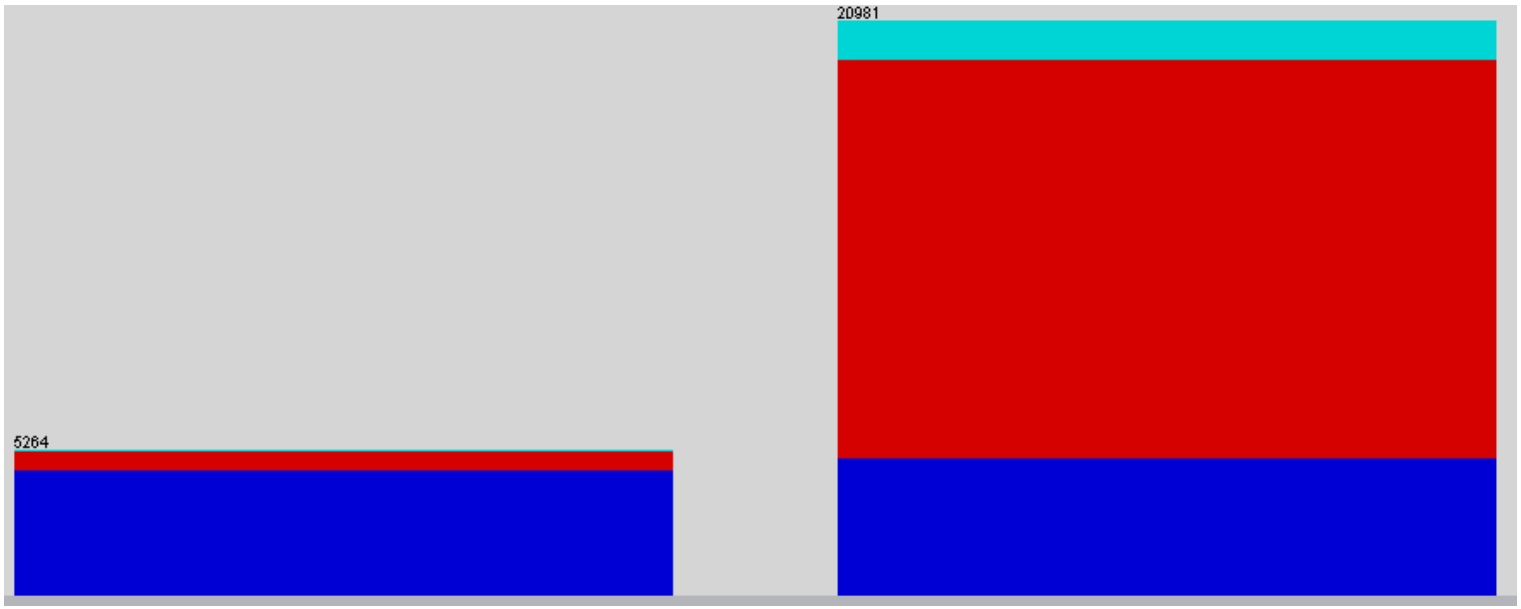

Figure 4.33: Baseline Osteoarthritis Distribution among Low Risk, Medium Risk, and High Risk Patients

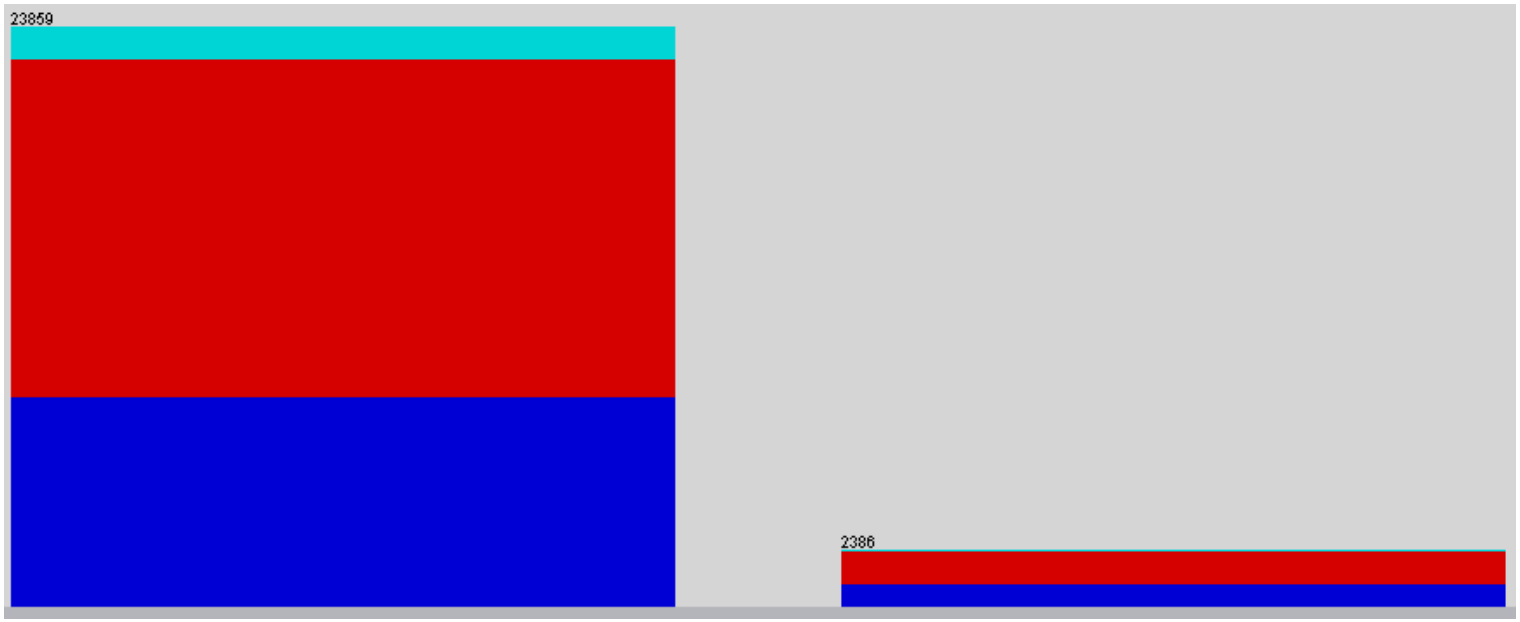

Figure 4.34: Baseline Psychiatric Disorder Distribution among Low Risk, Medium Risk, and High Risk Patients

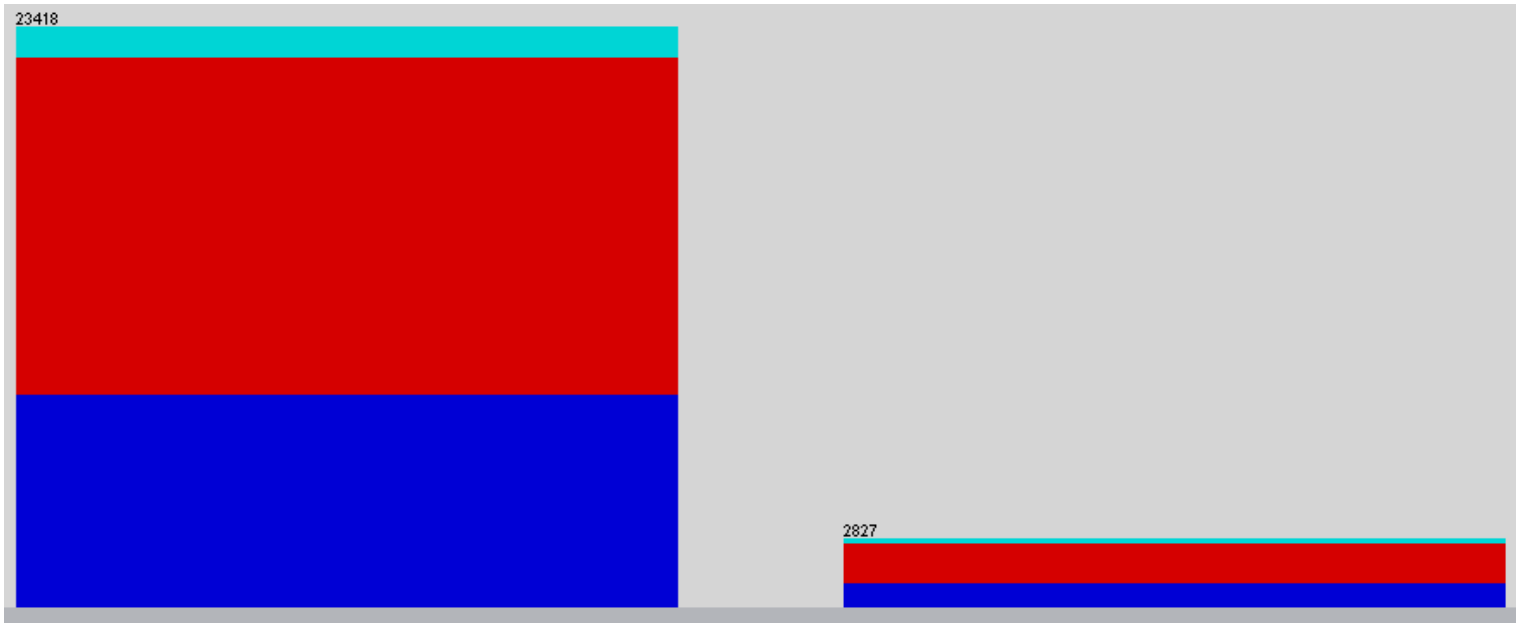

Figure 4.35: Baseline Pulmonary Disease Distribution among Low Risk, Medium Risk, and High Risk Patients

Table 4.33: Low Risk, Medium Risk, and High Risk Patients Data Characteristics - Baseline Charlson Index, Overall Incidence of 30-day Mortality for Each Surgery, and Overall Incidence of In-hospital Complications for Each Surgery

Baseline Charlson	Min Val. = 0, Max Val. = 13, Mean = 1.079, St. Dev. = 1.825, Missing Val. = 0, Distinct Val. = 14, Unique Val. = 0
ccsMort30rate	Min Val. = 0, Max Val. = 0.017, Mean = 0.004, St. Dev = 0.004, Missing Val. = 0, Distinct Val. = 21, Unique Val = 0
ccsComplicationRate	Min Val. = 0.016, Max Val. = 0.466, Mean = 0.13, St. Dev. = 0.087, Missing Val. = 0, Distinct Val. = 23, Unique Val. = 0

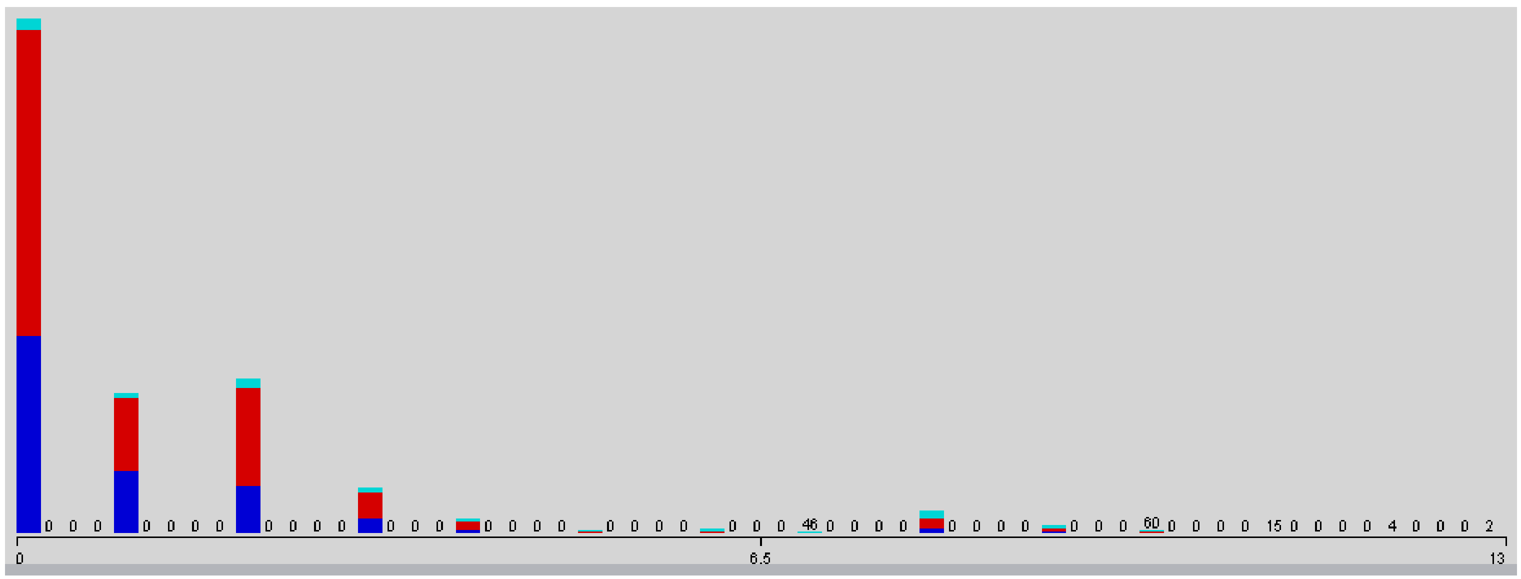

Figure 4.36: Baseline Charlson Index Distribution among Low Risk, Medium Risk, and High Risk Patients

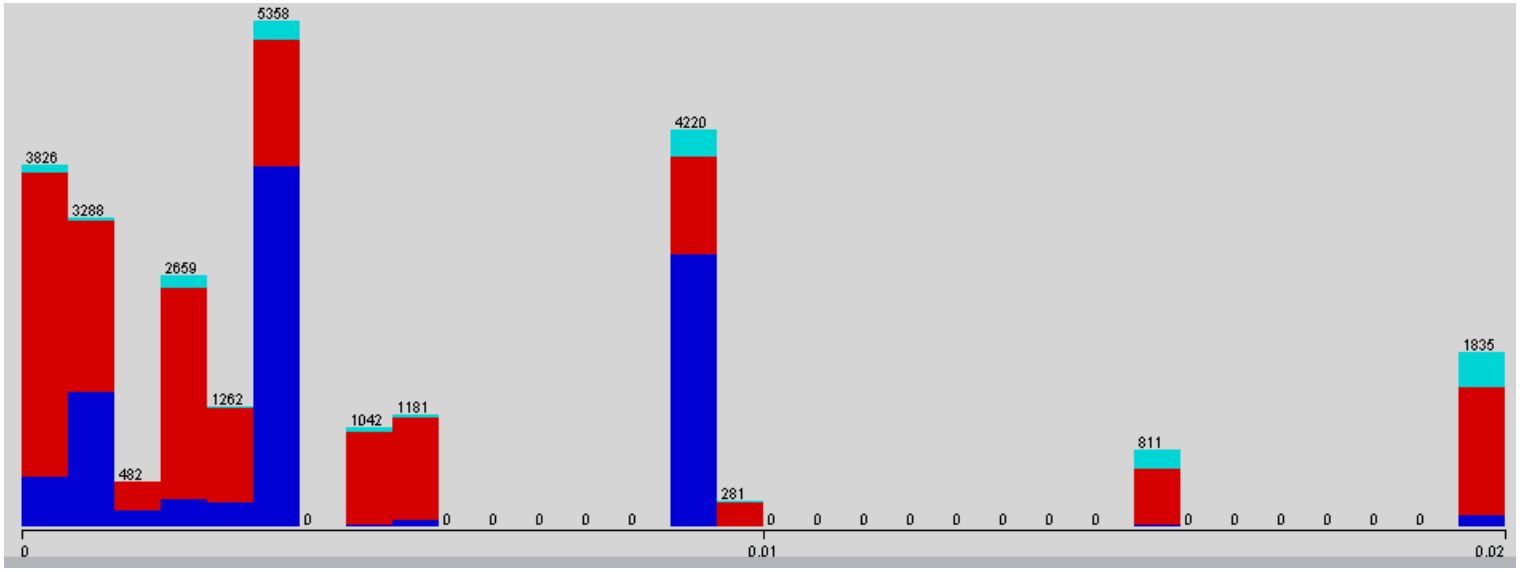

Figure 4.37: Overall Incidence of 30-day Mortality Distribution Low Risk, Medium Risk, and High Risk Patients

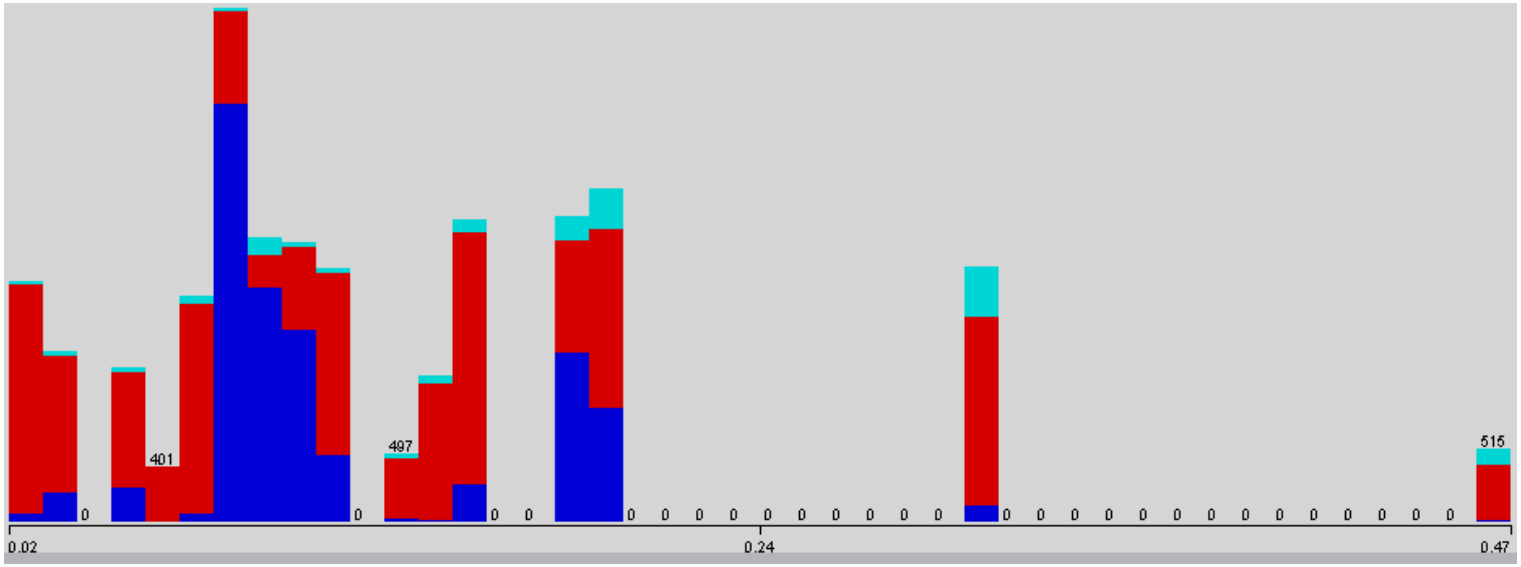

Figure 4.38: Overall Incidence of Complication Distribution Low Risk, Medium Risk, and High Risk Patients

Table 4.34: Low Risk, Medium Risk, and High Risk Patients Data Characteristics - Hour, Day, Month, Moon Phase of Surgery, 30-Mortality of Patients, In-hospital Complication of Patients

Hour	Min Val. = 6, Max Val. =19, Mean = 10.367, St. Dev. = 2.909, Missing Val. = 0, Distinct Val.= 767, Unique Val. = 27
Day of Week	
Monday	5,781
Tuesday	5,779
Wednesday	5,073
Thursday	4,631
Friday	4,981
Month	
January	2,171
February	2,039
March	2,245
April	2,255
May	2,156
June	2,452
July	1,884
August	2,606
September	2,656
October	2,218
November	2,075
December	1,488
Moon Phase	
First Quarter	6,698
Last Quarter	6,671
New Moon	6,321
Full Moon	6,555
mort30	
No	26,122
Yes	123
Complication	
No	22,990
Yes	3,255

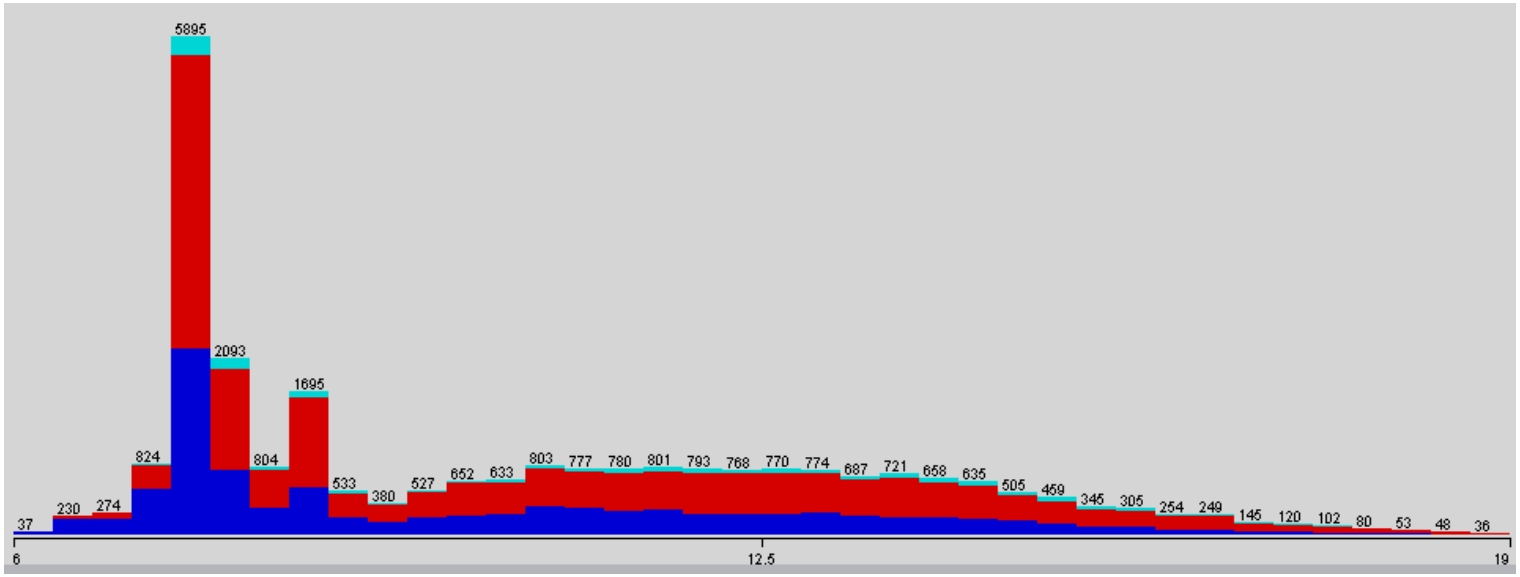

Figure 4.39: Hour of Surgery Distribution among Low Risk, Medium Risk, and High Risk Patients

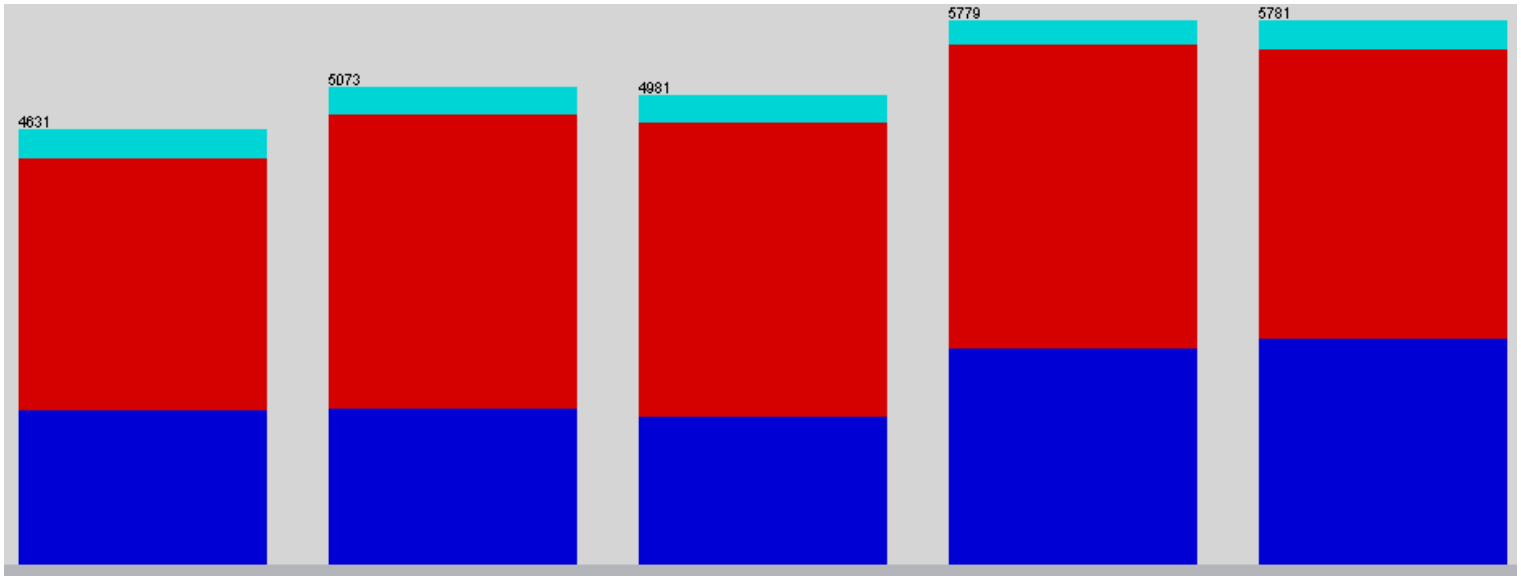

Figure 4.40: Day of Surgery Distribution among Low Risk, Medium Risk, and High Risk Patients

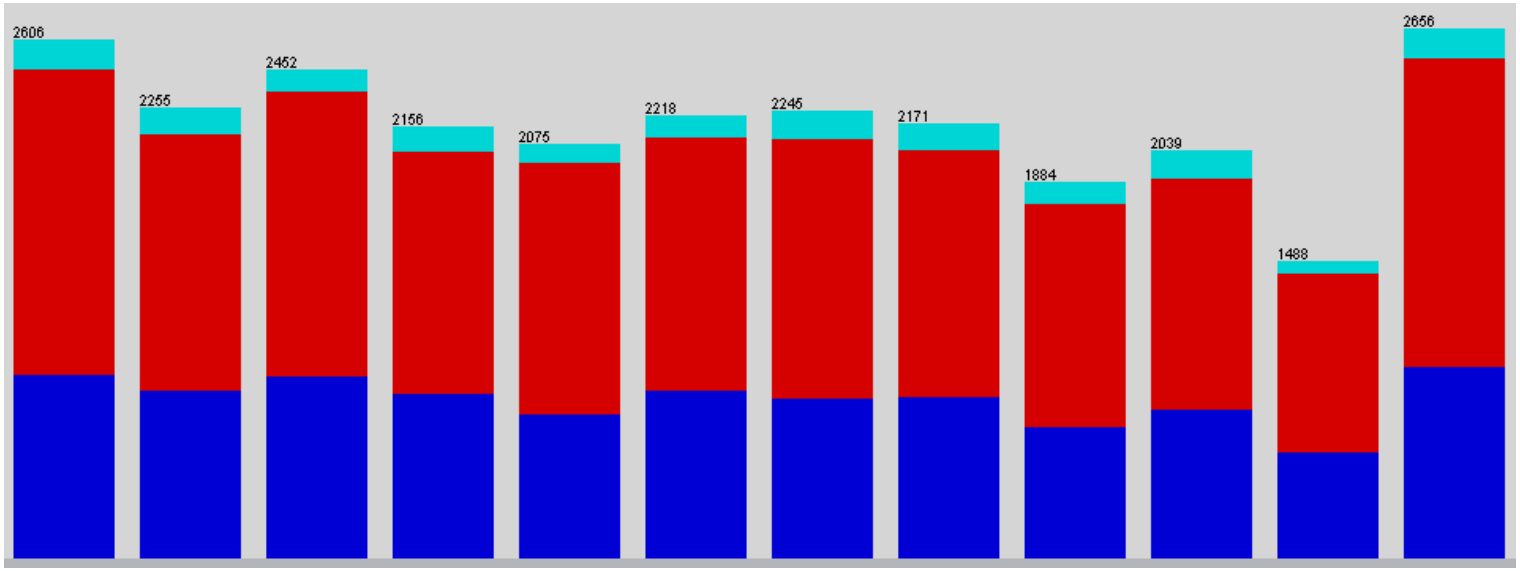

Figure 4.41: Month of Surgery Distribution among Low Risk, Medium Risk, and High Risk Patients

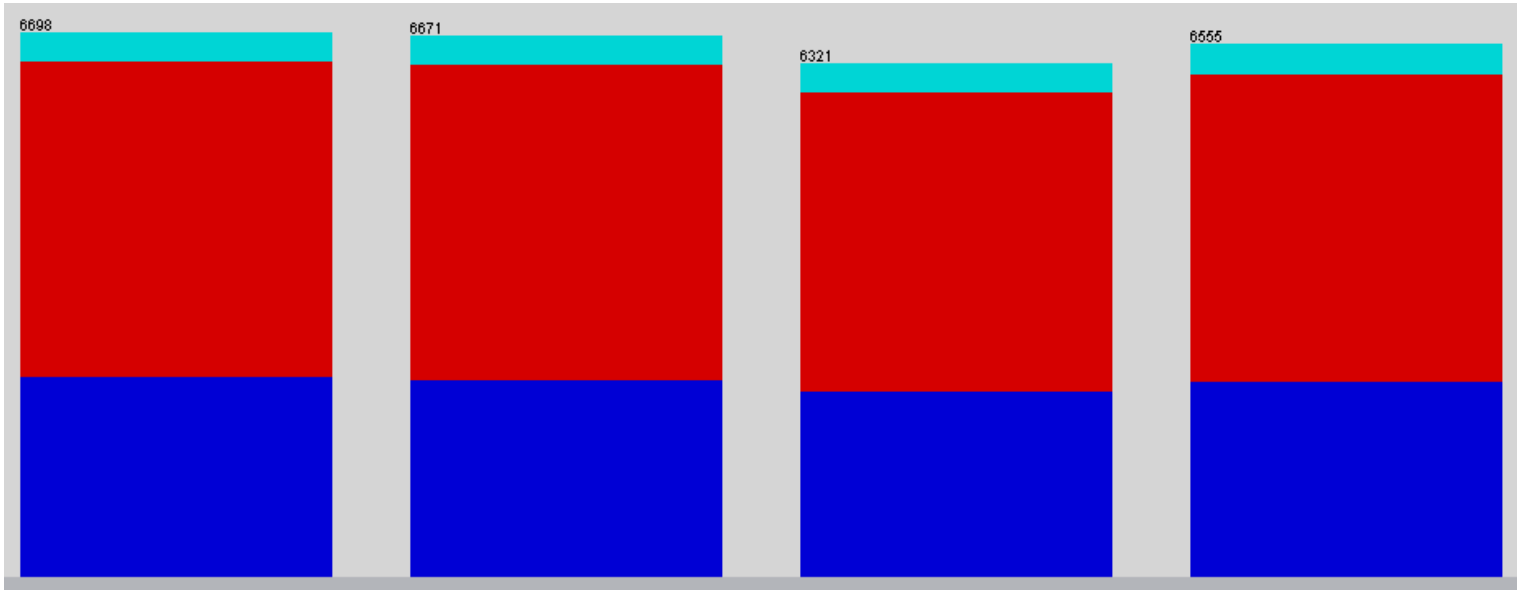

Figure 4.42: Moon Phase of Surgery Distribution among Low Risk, Medium Risk, and High Risk Patients

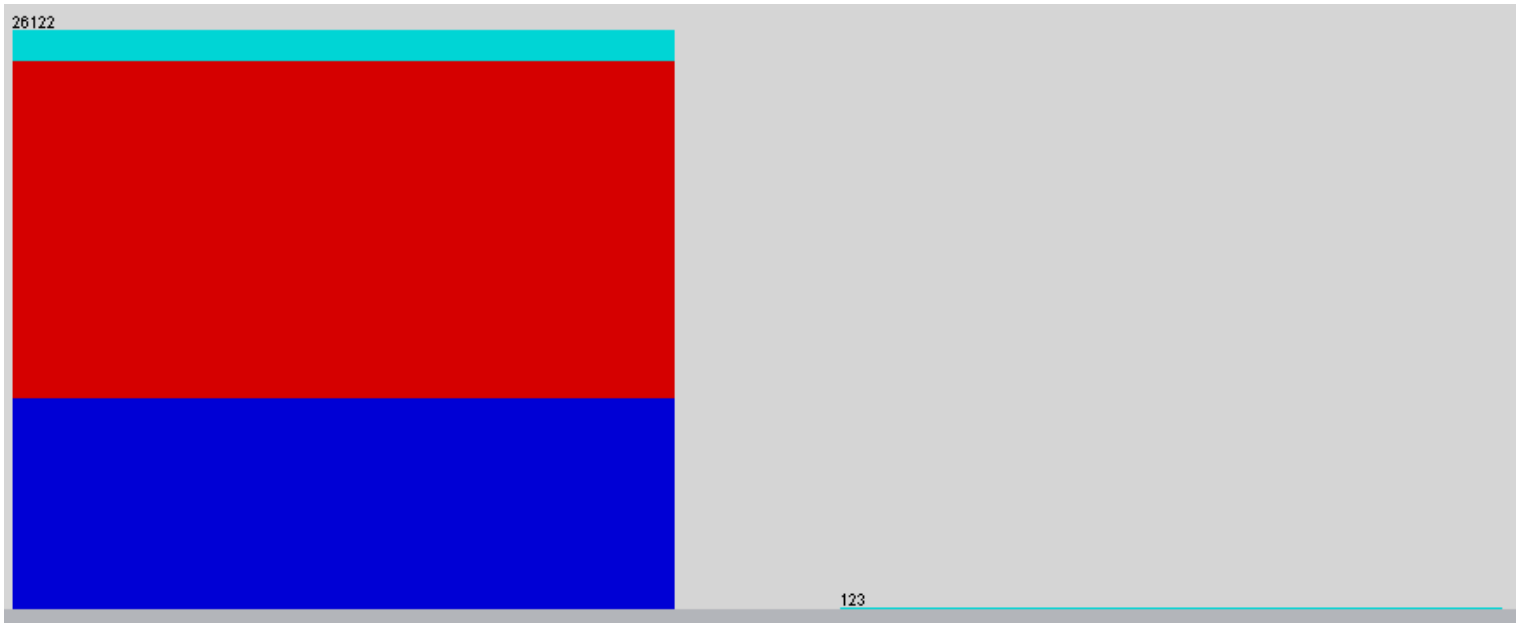

Figure 4.43: 30-day Patient Mortality Distribution among Low Risk, Medium Risk, and High Risk Patients

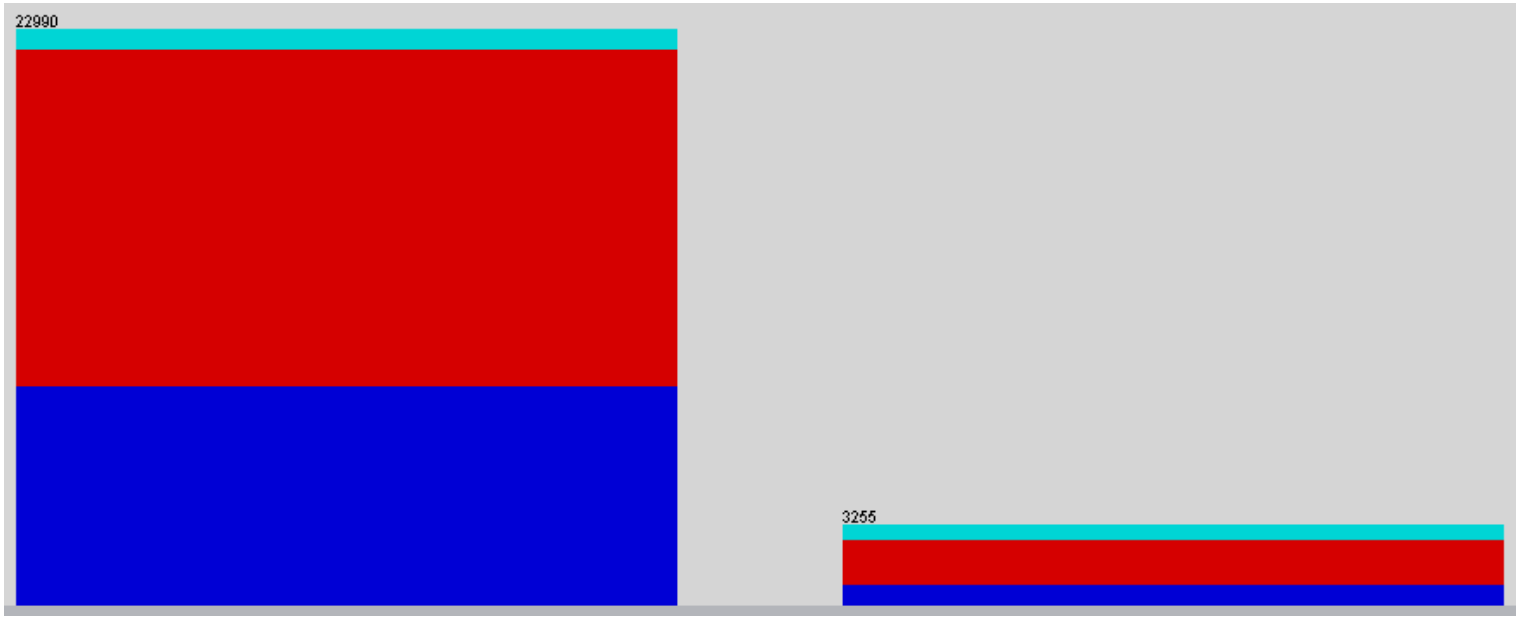

Figure 4.44: Patient Complication Distribution among Low Risk, Medium Risk, and High Risk Patients

4.2.2 Prediction of Low, Medium, and High Risk Patients

To predict low risk, medium risk, and high risk patients in "Surgery Timing LMH" dataset, we use 10×10-folding cross-validation experiments on seven commonly used and well-known classification methods, including Random Forest, Decision Trees, Nearest Neighbor, Logistic Regression, Naïve Bayes, Bayes Network, and Neural Networks [13]. Surgery Timing LMH dataset is randomly partitioned into ten approximately equal parts; one of these subsets is designated as "test set", a model is built on the remaining nine subsets which form the "training dataset", and then tested by predicting the classes of patients in the test set using a classification method. This procedure is repeated 10 times, always taking another one of the ten parts in the role of the test set (re-randomizing the patients into 10 new subsets and repeat the procedure 9 additional times) for a total of 100 tests for each of the nine classification methods. Tables 4.35-4.41 show the average accuracy, proportion of correctly classified low risk patients, proportion of correctly classified high risk patients as well as average precision, recall, F-measure (weighted mean of the precision and recall), and area under Receiver Operating Characteristic (ROC) curve for Random Forest, Decision Trees, Nearest Neighbor, Logistic Regression, Naïve Bayes, Bayes Network, and Neural Networks, respectively. For Surgery Timing LMH dataset, we were unable to obtain results using Stochastic Gradient method and Support Vector Machines.

Table 4.35: Cross-Validation of Low Risk, Medium Risk, and High Risk Patients Using Random Forest

Random Forest	**Average Cross-Validation Results**
Training Instances	23620.4
Testing Instance	2624.5
Number Correct	2297.4
Number Incorrect	327.1
Percent Correct	87.5%
Percent Incorrect	12.5%
Mean Absolute Error	0.196
Area Under ROC	0.714
F-Measure	0.933
True Positive Rate	0.989
Number of True Positives	2274.1
False Positive Rate	0.928
Number of False Positives	302.2
True Negative Rate	0.072
Number of True Negatives	23.3
False Negative Rate	0.011
Number of False Negatives	24.9
Weighted True Positive Rate	0.875
Weighted False Positive Rate	0.815
Weighted True Negative Rate	0.185
Weighted False Negative Rate	0.125
Weighted F-Measure	0.833
Weighted Area Under ROC	0.714

Table 4.36: Cross-Validation of Low Risk, Medium Risk, and High Risk Patients Using J48 Decision Tree

J48 Decision Tree	**Average Cross-Validation Results**
Training Instances	23620.4
Testing Instance	2624.5
Number Correct	2298.7
Number Incorrect	325.8
Percent Correct	87.6%
Percent Incorrect	12.4%
Mean Absolute Error	0.201
Area Under ROC	0.627
F-Measure	0.933
True Positive Rate	0.987
Number of True Positives	2270.6
False Positive Rate	0.914
Number of False Positives	297.4
True Negative Rate	0.086
Number of True Negatives	28.1
False Negative Rate	0.012
Number of False Negatives	28.4
Weighted True Positive Rate	0.875
Weighted False Positive Rate	0.802
Weighted True Negative Rate	0.198
Weighted False Negative Rate	0.124
Weighted F-Measure	0.835
Weighted Area Under ROC	0.627

Table 4.37: Cross-Validation of Low Risk, Medium Risk, and High Risk Patients Using k-Nearest Neighbor

k-Nearest Neighbor	**Average Cross-Validation Results**
Training Instances	23620.4
Testing Instance	2624.5
Number Correct	2125.9
Number Incorrect	498.6
Percent Correct	81.0%
Percent Incorrect	18.9%
Mean Absolute Error	0.189
Area Under ROC	0.555
F-Measure	0.892
True Positive Rate	0.894
Number of True Positives	2055.7
False Positive Rate	0.784
Number of False Positives	255.3
True Negative Rate	0.216
Number of True Negatives	70.2
False Negative Rate	0.106
Number of False Negatives	243.3
Weighted True Positive Rate	0.810
Weighted False Positive Rate	0.700
Weighted True Negative Rate	0.299
Weighted False Negative Rate	0.189
Weighted F-Measure	0.808
Weighted Area Under ROC	0.555

Table 4.38: Cross-Validation of Low Risk, Medium Risk, and High Risk Patients Using Logistic Regression

Logistic Regression	**Average Cross-Validation Results**
Training Instances	23620.4
Testing Instance	2624.5
Number Correct	2300.1
Number Incorrect	324.4
Percent Correct	87.6%
Percent Incorrect	12.4%
Mean Absolute Error	0.194
Area Under ROC	0.738
F-Measure	0.933
True Positive Rate	0.989
Number of True Positives	2275.9
False Positive Rate	0.926
Number of False Positives	301.4
True Negative Rate	0.074
Number of True Negatives	24.2
False Negative Rate	0.010
Number of False Negatives	23.0
Weighted True Positive Rate	0.876
Weighted False Positive Rate	0.812
Weighted True Negative Rate	0.187
Weighted False Negative Rate	0.123
Weighted F-Measure	0.833
Weighted Area Under ROC	0.738

Table 4.39: Cross-Validation of Low Risk, Medium Risk, and High Risk Patients Using Naïve Bayes

Naïve Bayes	**Average Cross-Validation Results**
Training Instances	23620.4
Testing Instance	2624.5
Number Correct	2177.6
Number Incorrect	446.9
Percent Correct	82.9%
Percent Incorrect	17.0%
Mean Absolute Error	0.193
Area Under ROC	0.728
F-Measure	0.902
True Positive Rate	0.899
Number of True Positives	2067.1
False Positive Rate	0.660
Number of False Positives	214.9
True Negative Rate	0.339
Number of True Negatives	110.5
False Negative Rate	0.101
Number of False Negatives	231.9
Weighted True Positive Rate	0.829
Weighted False Positive Rate	0.591
Weighted True Negative Rate	0.408
Weighted False Negative Rate	0.170
Weighted F-Measure	0.831
Weighted Area Under ROC	0.728

Table 4.40: Cross-Validation of Low Risk, Medium Risk, and High Risk Patients Using Bayes Network

Bayes Network	**Average Cross-Validation Results**
Training Instances	23620.4
Testing Instance	2624.5
Number Correct	2147.8
Number Incorrect	476.7
Percent Correct	81.8%
Percent Incorrect	18.2%
Mean Absolute Error	0.211
Area Under ROC	0.731
F-Measure	0.895
True Positive Rate	0.880
Number of True Positives	2023.5
False Positive Rate	0.617
Number of False Positives	201.2
True Negative Rate	0.382
Number of True Negatives	122.4
False Negative Rate	0.119
Number of False Negatives	275.5
Weighted True Positive Rate	0.818
Weighted False Positive Rate	0.556
Weighted True Negative Rate	0.443
Weighted False Negative Rate	0.181
Weighted F-Measure	0.826
Weighted Area Under ROC	0.731

Table 4.41: Cross-Validation of Low Risk, Medium Risk, and High Risk Patients Using Multi-layer Perceptron

Multi-layer Perceptron	**Average Cross-Validation Results**
Training Instances	23620.4
Testing Instance	2624.5
Number Correct	2218.7
Number Incorrect	405.8
Percent Correct	84.5%
Percent Incorrect	15.5%
Mean Absolute Error	0.160
Area Under ROC	0.651
F-Measure	0.914
True Positive Rate	0.938
Number of True Positives	2157.6
False Positive Rate	0.813
Number of False Positives	264.5
True Negative Rate	0.187
Number of True Negatives	61.0
False Negative Rate	0.061
Number of False Negatives	141.3
Weighted True Positive Rate	0.845
Weighted False Positive Rate	0.719
Weighted True Negative Rate	0.280
Weighted False Negative Rate	0.154
Weighted F-Measure	0.829
Weighted Area Under ROC	0.651

The average of 10×10-folding cross validation results for all seven classification methods are summarized in Table 4.42. The overall average accuracy of seven classification methods is 84.70%. Overall, the performance of the seven methods is validated by high values of prediction metrics: precision value of 0.89, recall value of 0.94, F-measure value of 0.91. The overall average value of area under ROC curves is 0.68.

Similar to the prediction of low and high risk mortality, we observe that all classification methods applied to Mortality RSI LMH dataset have comparable accuracy, precision, recall, F-measure and area under ROC curve. Logistic Regression provides the combination of best accuracy, precision, recall, and F-value as well as area under ROC curve.

Table 4.42: Average Cross-Validation Results for Seven Classfication Methods - Mortality RSI LMH Data

Classification Method	**Accuracy**	**Precision**	**Recall**	**F-Measure**	**Area under ROC Curve**
Random Forest	87.50	0.88	0.99	0.93	0.71
J48 Decision Tree	87.60	0.88	0.99	0.93	0.63
Nearest Neighbor	81.00	0.89	0.89	0.89	0.56
Logistic Regression	87.60	0.88	0.99	0.93	0.74
Naïve Bayes	82.90	0.91	0.90	0.90	0.73
Bayes Network	81.80	0.91	0.88	0.90	0.73
Multilayer Perceptron	84.50	0.89	0.94	0.91	0.65
Average	84.70	0.89	0.94	0.91	0.68

Chapter 5

Conclusion

In this thesis we integrate fundamental concepts from conventional statistics with the more explanatory, algorithmic, and computational techniques offered by machine learning to predict early mortality risk of surgical patients. Well-known and commonly used classification methods, including Random Forest, Decision Trees, Nearest Neighbor, Stochastic Gradient Descent, Logistic Regression, Naïve Bayes, Bayes Network, Neural Networks, and Support Vector Machines, are applied to predict low-risk, medium-risk, and high-risk mortality of elective general surgical patients treated between January 2005 and September 2010 at the Cleveland Clinic [33]. The mortality risk prediction is based on clinical factors including surgery type, age, gender, race, BMI, underlying chronic conditions, surgical risk indices, surgical timing predictors such as hour, day of week, month, moon phase as well as the 30-day mortality and in-hospital complication for each patient. We perform 10×10-folding cross validation experiments to evaluate the prediction performance of the classification methods on low, medium, and high mortality risk groups. The overall average accuracy of the classification methods applied to predict low-risk

and high-risk mortality is 85.20% with precision value of 0.89, recall value of 0.95, and F-measure value of 0.92. The overall accuracy of the classification method applied to predict low-risk, medium-risk, and high-risk mortality is 84.70% with precision value of 0.89, recall value of 0.94, and F-measure value of 0.91. A Decision Tree classification model consisting of 83 low risk patterns and 135 high risk patterns are presented to provide medical experts with an explainable classification model that can serve for further investigation of the clinical features associated with early mortality risk of surgical patients.

Bibliography

[1] S. Amari. Backpropagation and stochastic gradient descent method. *Neurocomputing*, 5(4-5):185–196, 1993.

[2] J.F. Avila-Herrera and M.M. Subasi. Logical analysis of multi-class data. In *2015 Latin American Computing Conference (CLEI)*, pages 1–10. IEEE, 2015.

[3] N. Aylein. Support vector machines: A simple explaination. *KD Nuggets*, 1:1–5, 2016.

[4] T.C. Bain, J.F. Avila-Herrera, E. Subasi, and M.M. Subasi. Logical analysis of multiclass data with relaxed patterns. *Annals of Operations Research*, 287(1):11–35, 2020.

[5] C.M. Bishop. *Pattern recognition and machine learning*. springer, 2006.

[6] P.C. Bonasso, M.S. Dassinger, M.L. Ryan, M.S. Gowen, J.M. Burford, and S.D. Smith. 24-hour and 30-day perioperative mortality in pediatric surgery. *Journal of pediatric surgery*, 54(4):628–630, 2019.

[7] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[8] C.J.C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[9] R.O. Duda, P.E. Hart, et al. *Pattern classification*. John Wiley & Sons, 2006.

[10] L.V. Fausett. *Fundamentals of neural networks: architectures, algorithms, and applications*. Prentice-Hall Englewood Cliffs, NJ, 1994.

[11] K. Foote. A brief history of machine learning. *Dataversity*, 1:1–7, 2019.

[12] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine learning*, 29(2):131–163, 1997.

[13] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutremann, and I. Witten. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1), 2009.

[14] F. Harrell. Why is logistic regression called a machine learning algorithm? *Stack Exchange*, 1:1–11, 2015.

[15] M. Harrison. *Machine Learning Pocket Reference Working with Structured Data in Python*. O'Reilly, 2019.

[16] O. Harrison. Machine learning basics with the k-nearest neighbors algorithm. *Towards Data Science*, 1:1–12, 2018.

[17] M. Horny. Bayesian networks. *Boston University School of Public Health*, 5:1–15, 2014.

[18] N. Kain. Understanding of multilayer perceptrons(mlp). *Medium*, 1:1–6, 2018.

[19] J. Kelleher and B. Tierney. *Data Science*. Massachusetts Institute of Technology, 2018.

[20] A.L. Luryi, M.M. Chen, S. Mehra, S.A. Roman, J.A. Sosa, and B.L. Judson. Hospital readmission and 30-day mortality after surgery for oral cavity cancer: Analysis of 21,681 cases. *Head & neck*, 38(S1):E221–E226, 2016.

[21] E.J. Mascha, D. Yang, S. Weiss, and D.I. Sessler. Intraoperative mean arterial pressure variability and 30-day mortality in patients having noncardiac surgery. *Anesthesiology*, 123(1):79–91, 2015.

[22] Ò. Miró, X. Rossello, V. Gil, F.J. Martín-Sánchez, P. Llorens, P. Herrero-Puente, J. Jacob, H. Bueno, and S.J. Pocock. Predicting 30-day mortality for patients with acute heart failure in the emergency department: a cohort study. *Annals of internal medicine*, 167(10):698–705, 2017.

[23] H.B. Mitchell and P.A. Schaefer. A "soft" k-nearest neighbor voting scheme. *International journal of intelligent systems*, 16(4):459–468, 2001.

[24] K.P. Murphy et al. Naive bayes classifiers. *University of British Columbia*, 18(60), 2006.

[25] F. Pedregosa. The stochastic gradient method. *fa.bianp.net*, 1:1, 2018.

[26] M. Pelillo. Who invented the nearest neighbor rule? *37 Steps*, 1:1, 2014.

[27] J. Platt. Sequential minimal optimization: A fast algorithm for training support vector machines. *Microsoft Research*, 1:1–21, 1998.

[28] B. Poczos. Introduction to machine learning perceptron. *Carnegie Mellon University*, 1:1–10, 2019.

[29] J.R. Quinlan. *C4. 5: programs for machine learning*. Elsevier, 2014.

[30] N. Saravann and V. Gayathri. Performance and classification evaluation of j48 algorithm and kendall's based j48 algorithm(knj48). *International Journal of Computational Intelligence and Informatics*, 7:188–198, 2018.

[31] B. Schölkopf and A. J. Smola. *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. The MIT Press, Cambrige, MA, 2001.

[32] C. Sehra. Decision trees explained easily. *Medium*, 1:1–6, 2018.

[33] D.I. Sessler, A. Kurz, L. Saager, and J.E. Dalton. Operation timing and 30-day mortality after elective general surgery. *Anesthesia & Analgesia*, 113(6):1423–1428, 2011.

[34] M. Simard, C. Sirois, and B. Candas. Validation of the combined comorbidity index of charlson and elixhauser to predict 30-day mortality across icd-9 and icd-10. *Medical care*, 56(5):441–447, 2018.

[35] N. Singh-Miller and M. Collins. Learning label embeddings for nearest-neighbor multi-class classification with an application to speech recognition. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1678–1686. 2009.

[36] A. Srinivason. Stochastic gradient descent - clearly explained!! *Towards Data Science*, 1:1–5, 2019.

[37] A. Tandel. Support vector machines - a brief overview. *Towards Data Science*, 1:1, 2017.

[38] Z. Tang, K. Yang, M. Zhong, R. Yang, J. Zhang, Q. Jiang, and H. Liu. Predictors of 30-day mortality in traumatic brain-injured patients after primary decompressive craniectomy. *World neurosurgery*, 134:e298–e305, 2020.

[39] P. Vadapalli. Naive bayes explained: Function, advantages & disadvantages, application in 2020. *upGrad blog*, 1:1–14, 2020.

[40] A. Yadav. Support vector machines(svm). *Towards Data Science*, 1:1–17, 2018.

[41] T. Yiu. Understanding random forest. *Towards Data Science*, 1:1–11, 2019.

[42] E. Zhang and Y. Zhang. F-measure. *Encyclopedia of Database Systems*, 1:50, 2009.

[43] J. Zornoza. Logistic regression explained. *Towards Data Science*, 1:1–8, 2020.