Cancer prognosis & Survival rate prediction using machine learning algorithm.

Submitted by-

Borhan Uddin Alif
T181020

A thesis submitted to the Department of Electronics and Telecommunication Engineering in partial fulfillment of the requirements for the degree of
B.Sc. in ETE.

Department of Electronics and Telecommunication Engineering,
International Islamic University Chittagong.

Declaration

It is hereby declared that

1. The thesis we submitted for our IIUC degree is our unique work.

2. The thesis does not include any content that has been published before or that was written by a third party unless it is properly cited with complete and correct referencing.

3. The thesis does not include any content that has been approved or submitted for consideration for any other degree or certificate at a university or other institution.

4. We have given credit to all major sources of assistance.

Student's Full Name & Signature:

Borhan Uddin Alif (T181020)

Approval

The thesis titled “Cancer Detection Using Machine Learning” is submitted by

Miftahul Jannah (T181024)

Of 18th batch of ETE has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in ETE in 2022.

Examining Committee:

Supervisor:
(Member)

Md. Ibrahim Rupam
Assistant Professor
Dept. of ETE, IIUC.

Program Coordinator:
(Member)

Eng. Abdul Gafur
Associate Professor
Dept. of ETE, IIUC.

Head of Department:
(Chair)

Eng. Syed Zahidur Rashid
Chairman,
Dept. of ETE, IIUC.

Abstract

Roughly 10 million deaths, or nearly one in six deaths, were caused by cancer in 2020, making it the top cause of death globally. Breast, lung, colon, rectum, and prostate cancers are the most prevalent types of cancer. Any disease that can affect any region of the body is referred to as cancer. Neoplasms and malignant tumors are other words that are used. One characteristic of cancer is the quick development of aberrant cells that expand outside of their normal borders, infiltrate other body components, and eventually move to other organs. This process is known as metastasis. The main reason why cancer patients die is because of widespread metastases [1].

The research was held on two different datasets of breast cancer. To train our model initially we used a primary data set from UCI which includes 569 instances with several attributes [2]. This was for cancer diagnosis prediction/detection. Then we used another dataset that includes 4024 instances with 14 different attributes which we mainly used for death rate prediction with total analysis [3]. Here 6 algorithms were incorporated including KNN, Random Forest, J48, etc., and their results were compared.

Keywords: Cancer Diagnosis, Breast Cancer, KNN, Random Forest, Feature Selection

Dedication

This study is intended to help people with different types of cancer seeking to find a cure or make a decision based on any prediction model.

Acknowledgment

First, importantly, we want to express our gratitude to Allah for His unwavering assistance in enabling us to carry on with our research devoid of significant obstacles. Additionally, we wanted to express our gratitude to the supportive professors and, in particular, to our supervisor for putting up with our errors and providing ongoing criticism to help us improve our work. We also want to express our gratitude to our parents and friends for their unwavering support throughout the semester.

Table of Contents

Figures

List of Tables

Chapter 1 Introduction

1.1 Introduction

Any disease that can affect any region of the body is referred to as cancer. Neoplasms and malignant tumors are other words that are used. One characteristic of cancer is the quick development of aberrant cells that expand outside of their normal borders, infiltrate other body components, and eventually move to other organs. This process is known as metastasis. The main reason why cancer patients die is because of widespread metastases. According to research by WHO in 2020 the death's number of cancer will be increased to 10 million in a year. And in the stated year the mostly affected cancer is Breast cancer. Almost 2.26 million people are affected by breast cancer. That's why we decided to do this research on breast cancer [4].

Cancer has been described as a diverse illness with a wide range of subgroups. Early cancer diagnosis and prognosis are essential for clinical patient treatment, which has become a requirement in cancer research. Numerous research teams from the biomedical and bioinformatics fields have studied the use of machine learning (ML) techniques due to the significance of categorizing cancer patients into high or low-risk groups. These methods have been applied to simulate the development and management of malignant diseases. Furthermore, their significance is demonstrated by the fact that ML algorithms can recognize important features in complicated datasets. Several of these methods, such as Support Vector Machines (SVMs), Artificial Neural Networks (ANNs), Bayesian Networks (BNs), and Decision Trees (DTs) have been widely applied in cancer research for the development of predictive models, resulting in effective and accurate decision making. Although it is clear techniques can enhance our comprehension of how cancer progresses, more validation is required before these techniques can be used in routine clinical practice. We cover contemporary ML techniques used in the simulation of cancer progression in this paper. The presented prediction models are built using a variety of supervised ML techniques, as well as numerous input attributes and data samples [5].

Our main vision was to find out a prediction model which can predict or detect cancer with a tolerable percentage of accuracy. Our thesis contains and works on two different datasets with several attributes. Through the feature engineering process, we select the attributes which a big factor in case of the diagnosis of the can Here are our three main objectives to do this work are of particular interest in cancer prognosis and prediction:

1) The risk assessment of cancer susceptibility; 2) the forecasting of cancer recurrence; and 3) the forecasting of cancer survival

1.2 Problem Statement

14 million new patients with cancer around the world. That's millions of people who'll face years of uncertainty. Pathologists have been performing cancer diagnoses and prognoses for decades. Most pathologists have a 96–98% success rate for diagnosing cancer. They're pretty good at that part. The problem comes in the next part. According to the Oslo University Hospital, the accuracy of prognoses is only 60% for pathologists. A prognosis is the part of a biopsy that comes after cancer has been diagnosed, it is predicting the development of the disease. This is where machine learning steps in. ML has key advantages over Pathologists. Besides, if cancer diagnosis and the chance of survival can be predicted, the treatment process gets quite easier and more specific [6].

Cancer research has long used machine learning. Since almost 20 years ago (Simes 1985; Maclin et al. 1991; Ciccheti 1992), artificial neural networks (ANNs) and decision trees (DTs) have been employed in the detection and diagnosis of cancer. Today, a wide variety of applications apply machine learning techniques, from the detection and classification of tumors using X-ray and CRT images to the classification of malignancies from proteomic and genomic (microarray) assays (Petricoin and Liotta 2004; Bocchi et al. 2004). (Zhou et al. 2004; Dettling 2004; Wang et al. 2005). We wanted to establish a machine learning-based model with which we can easily identify a person's stage of breast cancer. This stage may indicate that this person needs help [7].

In this field, different research paper offers different models or algorithms. We find the gap in accuracy in some recent research on breast cancer prognosis and survival rate prediction. We will work on the accuracy rate with proper feature selection and machine learning algorithms [8] [9] [10] [11].

1.3 Aim of Study

Our aim with this study is to find a Machine Learning based algorithm that can predict cancer and predict survival rate or death rate accurately. Here are our three main objectives to do this work are of particular interest in cancer prognosis and prediction:

- The risk assessment of cancer susceptibility;
- The forecasting of cancer recurrence; and
- The forecasting of cancer survival

1.4 Research Methodology

We go through a well-decorated process to complete our thesis. Selecting the subject for the thesis to submit the final thesis is part of the methodology. After determining the subject for the thesis we set up a plan to finish our work properly. Though we changed some of our steps according to necessity didn't forget our final goal. The following flow-chart will give you an idea about our research methodology:

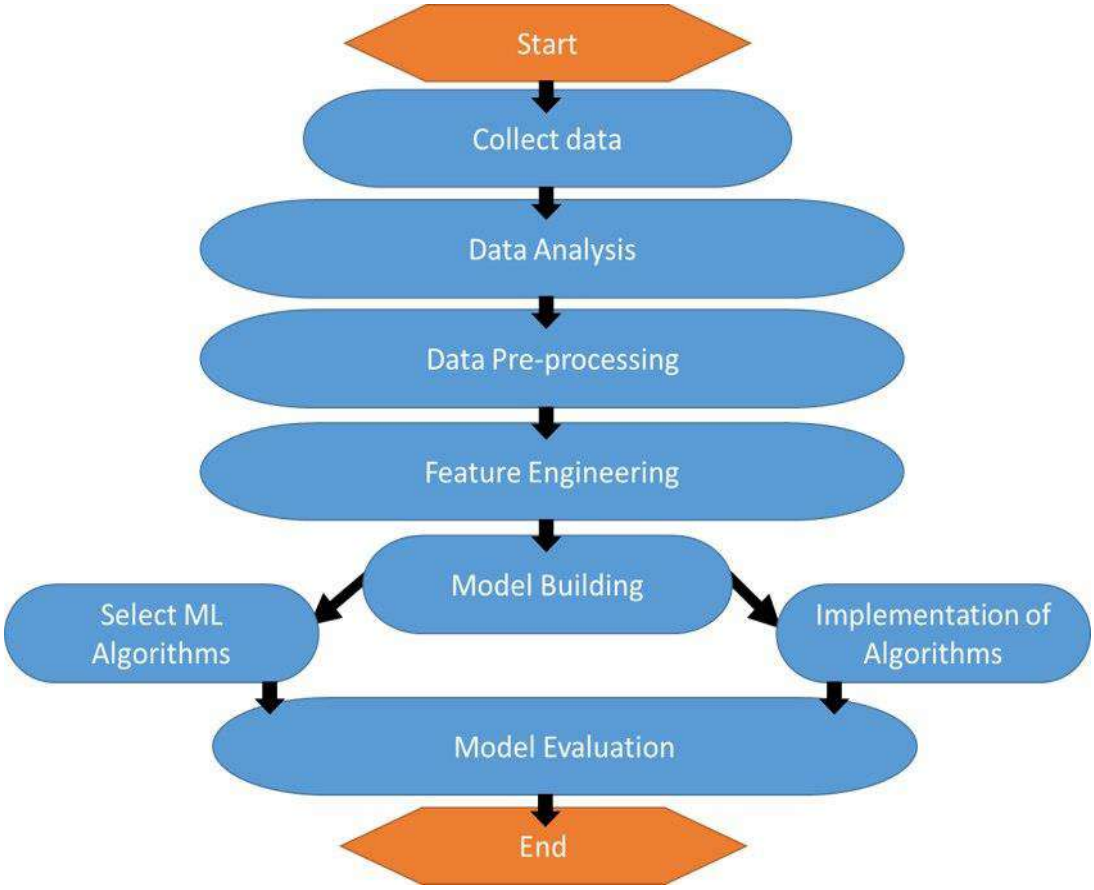


Figure 1: Flow chart of Research Methodology

1.5 Thesis Outline

This report puts influences constructing a prediction model which would be beneficial in detecting a potential addict in the primary stage. The authors aim to formulate a dataset from the context of our country which would be used for training existing supervised machine learning models to classify new observations. The overall report focuses on the steps that were followed by the researchers.

Firstly, the introduction part (Chapter 1) states the motivation behind the research that inspired authors to address this particular problem statement. The goals of our research and a summary of the work are briefly discussed here.

The chapter named “Theoretical Background” (Chapter 2) will give you a better understanding of every important term used in this thesis book or research process.

In the Literature review section (Chapter 3), we have discussed papers from the same background which have addressed a similar issue. In addition to that, some statistical and psychological papers are mentioned which refer to the available secondary data. The purpose of the background study was to find out the shortcomings of previous research.

In the Methodology section Chapter 44), we have discussed our planning to carry out our research. This section is about a logical, systematic plan to resolve the research problem we choose. What data we’re going to collect and where from, as well as how it's being collected and analyzed? All these questions will be answered in this section.

In the result analysis phase (Chapter 5), includes our proposed models and a comparative study of the prediction rate among respective models/algorithms. As we used two datasets for this research we will analyze both of the datasets.

Chapter 6 concludes the thesis with the future scope of this thesis.

Chapter 2 Theoretical Background

2.1 Breast Cancer

Breast cancer is a condition in which the breast's cells proliferate out of control. Breast cancer comes in several forms. Which breast cells develop into cancer determines the type of breast cancer.

Different areas of the breast might give rise to breast cancer. There are three basic components of a breast: connective tissue, ducts, and lobules. The glands that generate milk are called lobules. Milk travels through tubes called ducts to the nipple. The connective tissue, which is made up of fatty and fibrous tissue, envelops and holds everything in place. The ducts or lobules are where most breast cancers start. Blood and lymph vessels are two ways that breast cancer can travel outside of the breast. Whenever breast cancer spreads to a different body it is said to have metastasized [12].

Kinds of Breast Cancer

The most common kinds of breast cancer are—

- **Invasive ductal carcinoma.** The cancer cells begin in the ducts and then grow outside the ducts into other parts of the breast tissue. Invasive cancer cells can also spread, or metastasize, to other parts of the body.

- **Invasive lobular carcinoma.** Cancer cells begin in the lobules and then spread from the lobules to the breast tissues that are close by. These invasive cancer cells can also spread to other parts of the body.

There are several other less common kinds of breast cancer, such as Paget's disease, medullary, mucinous, and inflammatory breast cancer.

Ductal carcinoma *in situ* (DCIS) is a breast disease that may lead to invasive breast cancer. The cancer cells are only in the lining of the ducts and have not spread to other tissues in the breast [13].

Risk Factors of breast cancer

- **Getting older.** The risk for breast cancer increases with age. Most breast cancers are diagnosed after age 50.

- **Genetic mutations.** Women who have inherited changes (mutations) to certain genes, such as BRCA1 and BRCA2, are at higher risk of breast and ovarian cancer.

- **Reproductive history.** Starting menstrual periods before age 12 and starting menopause after age 55 expose women to hormones longer, raising their risk of getting breast cancer.

- **Having dense breasts.** Dense breasts have more connective tissue than fatty tissue, which can sometimes make it hard to see tumors on a mammogram. Women with dense breasts are more likely to get breast cancer.

- **Personal history of breast cancer or certain non-cancerous breast diseases.** Women who have had breast cancer are more likely to get breast cancer a second time. Some non-cancerous breast diseases such as atypical hyperplasia or lobular carcinoma *in situ* are associated with a higher risk of getting breast cancer.

- **Family history of breast or ovarian cancer.** A woman's risk for breast cancer is higher if she has a mother, sister, or daughter (first-degree relative) or multiple family members on either her mother's or father's side of the family who has had breast or ovarian cancer. Having a first-degree male relative with breast cancer also raises a woman's risk.

- **Previous treatment using radiation therapy.** Women who had radiation therapy to the chest or breasts (for instance, treatment of Hodgkin's lymphoma) before age 30 have a higher risk of getting breast cancer later in life.

- **Exposure to the drug diethylstilbestrol (DES).** DES was given to some pregnant women in the United States between 1940 and 1971 to prevent miscarriage. Women who took DES, or whose mothers took DES while pregnant with them, have a higher risk of getting breast cancer [14] [15].

2.2 Machine Learning

In the area of artificial intelligence known as machine learning (ML), statistical methods are applied to give computer systems the ability to gradually "learn" and develop on their own without being explicitly programmed. Machine learning investigates the study and development of algorithms that can pick up knowledge from data and datasets and generate predictions based on various teaching techniques. [16] Arthur Samuel first used the phrase "machine learning" in 1959 [17].

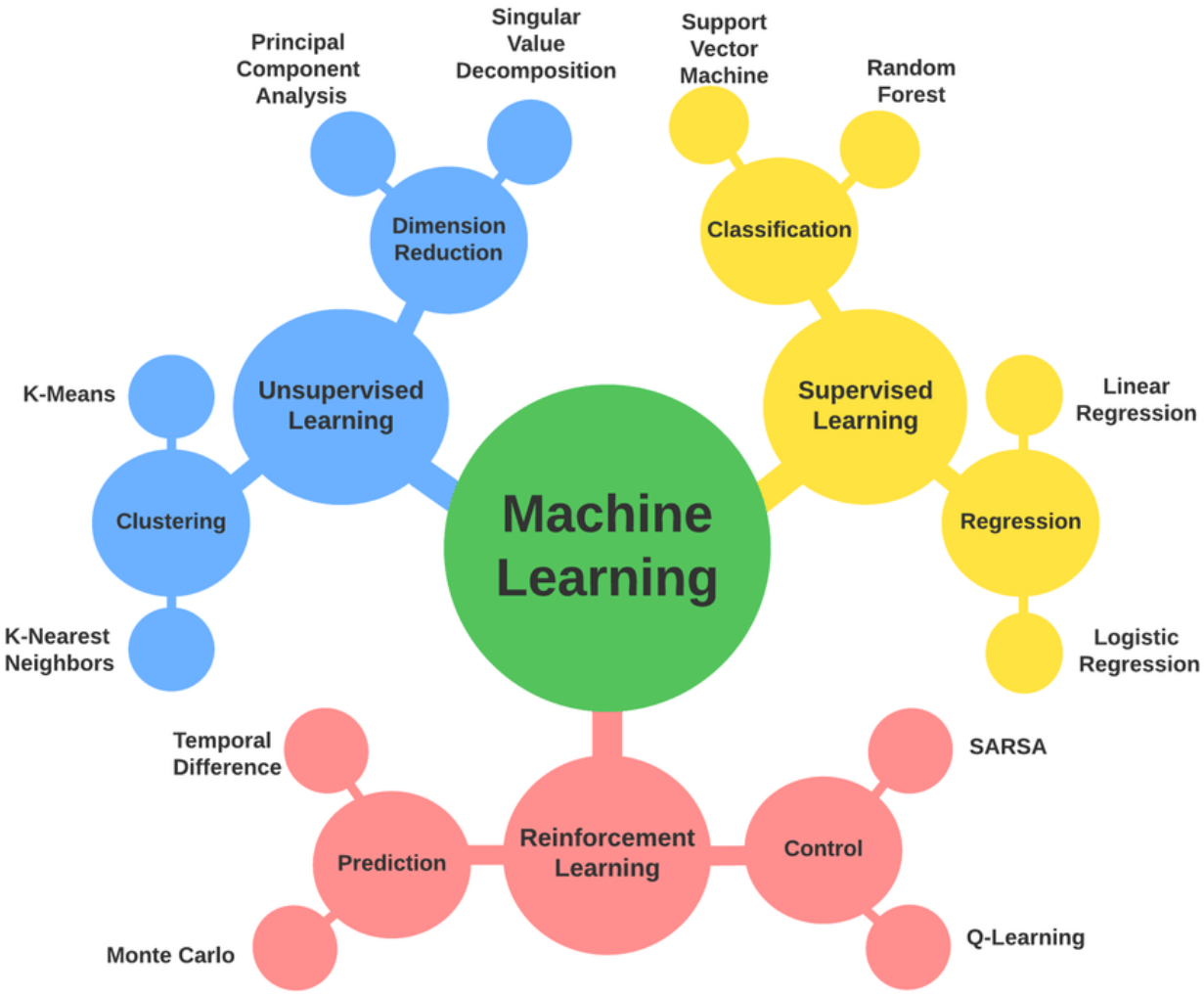


Figure 2: Machine Learning in details

The following three crucial machine learning categories are illustrated below:

1. **Supervised Learning:** In supervised learning, the computer uses labeled data—some of which have already been annotated with the right responses—to figure out how to transfer the input to the output. The objective is to estimate the mapping function as closely as possible so that you can forecast the output variables (Y) for new input data (x).
2. **Unsupervised Learning:** Neither classed nor labeled data are used to train the system. This enables the algorithm to carry out calculations independently and without supervision. The machine's job is to sort the unsorted data into groups based on similarities and differences to discover the hidden structure on its own.
3. **Reinforcement Learning:** The machine or the agent learns how to respond in a given environment by taking actions and observing the outcomes, which enables it to choose the best course of action in a given situation [18].

The tremendous amount of data that is currently available makes it impossible for humans to keep up with and analyze them. The primary focus of machine learning, a field of artificial intelligence and a subset of computer science, is on creating algorithms to solve this issue. Recent developments in this area have made a wide range of practically infinite applications possible, from data security to the financial and medical industries. However, there is still a lot of room for advancement in areas like social media services, disease identification, and prediction, virtual assistants, search engine optimization, fraud detection, manufacturing, etc. It will only get better and more seamlessly integrated into our daily lives, making life easier.

2.3 Algorithms

In this research, we used 6 different algorithms. We have to use different algorithms to compare the accuracy of every individual algorithm to get a clear view of our research.

Decision tree (j48): The foundation of J48 is a top-down, recursive divide-and-conquer method. The instances are divided into subsets, one for each branch that extends from the root node, after you choose which attribute to split on at the root node and create a branch for each conceivable attribute value [19].

KNN (K Nearest Neighbor): The supervised machine learning technique known as the k-nearest neighbors (KNN) can be used to tackle classification and

regression issues. We are frequently informed that you and your closest peers have a lot in common, whether it be your way of thinking, your professional conduct, your philosophical beliefs, or other aspects. As a result, we develop friendships with others whom we perceive to be like us. The same idea is used by the KNN algorithm. To determine what class a new unknown data point belongs to, it seeks to locate all of its nearest neighbors. It uses a distance-based strategy [20].

Logistic Regression: Based on prior observations, the statistical technique of logistic regression is used to forecast the outcome of a dependent variable. It is a common approach for tackling binary classification issues and is a subset of regression analysis. The most popular application of logistic regression, when the result is a binary choice, is binary logistic regression (yes or no) [21].

Support Vector Machine: SVM categorizes data points even when they are not otherwise linearly separable by mapping the data to a high-dimensional feature space. Once a separator between the categories is identified, the data are converted to enable the hyperplane representation of the separator. SVM categorizes data points even when they are not otherwise linearly separable by mapping the data to a high-dimensional feature space. Once a separator between the categories is identified, the data are converted to enable the hyperplane representation of the separator. The group to which a new record should belong can therefore be predicted using the features of new data [22].

Naïve Bayes: The Nave Bayes algorithm is a supervised learning method for classification issues that is based on the Bayes theorem. It is mostly employed in text categorization with a large training set. The Naive Bayes Classifier is one of the most straightforward and efficient classification algorithms available today. It aids in the development of quick machine-learning models capable of making accurate predictions. Being a probabilistic classifier, it makes predictions based on the likelihood that an object will occur [23].

Random forest: Supervised machine learning algorithms like random forest are frequently employed in classification and regression issues. On various samples, it constructs decision trees and uses their average for classification and majority vote for regression. The Random Forest Algorithm's ability to handle data sets with both continuous variables, as in regression, and categorical variables, as in classification, is one of its most crucial qualities. In terms of classification issues, it delivers superior outcomes [24].

2.4 Feature engineering

Feature engineering is the process of modifying your data set, including addition, deletion, combination, and mutation, to enhance the training of your machine learning model and achieve improved accuracy and performance. A solid understanding of the business issue and the data sources at hand is the foundation for effective feature engineering.

You gain a deeper grasp of your data and gain more insightful knowledge by developing new features. Feature engineering is one of the most useful data science approaches when done properly, but it is also one of the most difficult.

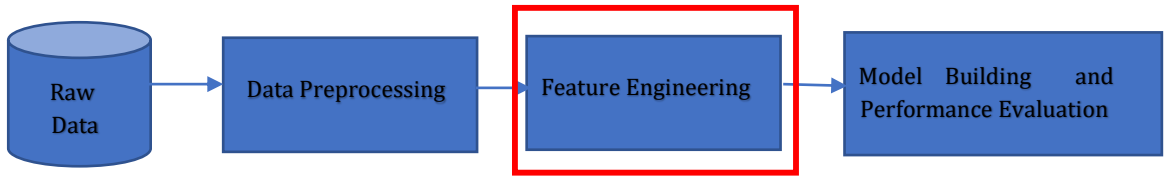


Figure 3: Feature Engineering

Some common types of feature engineering include:

- **Scaling and normalization** mean adjusting the range and center of data to ease learning and improve the interpretation of the results.
- **Filling missing values** implies filling in null values based on expert knowledge, heuristics, or some machine learning techniques. Real-world datasets can be missing values due to the difficulty of collecting complete datasets and because of errors in the data collection process.
- **Feature selection** means removing features because they are unimportant, redundant, or outright counterproductive to learning. Sometimes you simply have too many features and need fewer.
- **Feature coding** involves choosing a set of symbolic values to represent different categories. Concepts can be captured with a single column that comprises multiple values, or they can be captured with multiple columns, each of which represents a single value and has a true or false in each field. For example, feature coding can indicate whether a particular row of data was collected on a holiday. This is a form of feature construction.
- **Feature construction** creates a new feature(s) from one or more other features. For example, using the date you can add a feature that indicates the

day of the week. With this added insight, the algorithm could discover that certain outcomes are more likely on a Monday or a weekend.

- **Feature extraction** means moving from low-level features that are unsuitable for learning - practically speaking, you get poor testing results - to higher-level features that are useful for learning. Often feature extraction is valuable when you have specific data formats - like images or text - that has to be converted to a tabular row-column, example-feature format [25] [26].

Chapter 3 Related Work

With the development of technology, many contemporary methods for breast cancer prediction have emerged. The following is a brief description of the work in this field.

Some of the papers [27], [28] and [29]presented research related to disease prediction and diagnosis utilizing machine learning approaches, such as decision trees for cancer detection. According to Jini [30], the KNN method is one of the most often used classification algorithms in machine learning because of its well-known simplicity and adaptability in implementation.

A machine learning-based model for cancer diagnosis was put forth by Liu Lei [31]. The Sklearn machine learning library's Logistic Regression technique was utilized in this study to categorize the breast cancer data sets. The classification accuracy was 96.5% using the two criteria of maximum texture and minimal perimeter.

In another paper, they provided seven machine learning algorithms and an ontology model, as well as a comparison of their performance. Furthermore, two test modes are employed: 10-fold cross-validation and percentage split, and several performance measures such as Accuracy, F-Measure, Precision, and Recall are employed to assess the outcomes. The findings show that the ontological model has the uppermost accuracy even when no feature selection is used. This brings us to a new search area, to which we advise and urge academics to participate and produce new insights in the same context, to provide additional outcomes and analysis, to forecast, recommendations, or decisions, and so on. [32]

A paper by Melissa Zhao, Yushi Tang, Hyunkyung Kim, and Kohei Hasegawa shows a very interesting report to us. The analytic cohort is composed of 1874 patients with breast cancer. Overall, the median age was 62 years; the 5-year survival rate was 75%. ROC and accuracy were not significantly different between models (ROC and accuracy were around 0.67 and 0.72 across models, respectively). However, ensemble methods resulted in better fit (CS) with stable measures of variable importance across 10 random training/ validation splits. K-means clustering of gene expression profiles on training data points along with KNN classification of validation data points was a robust method of dimensional reduction. Furthermore, the gene expression cluster with the highest mortality risk was an influential factor in model prediction [33].

The researchers discovered that the Logistic Regression model had the highest accuracy in another paper. Artificial Neural Networks (ANN) displayed the highest

specificity, while Decision Trees (J48) displayed the maximum sensitivity. The bayesian model was generally more sensitive to death prediction, while the Decision Trees model was generally more sensitive to survival prediction.

Another paper by Dr. Shantosh Kumar and his team discovers the best highlights that are valuable for the depiction and analysis of breast cancer. It for the most part centers on the two acclaimed arrangement calculations i.e., K-NEAREST NEIGHBOUR order and Random Forest Classification Algorithms. The forecast and conclusion of breast cancer are particularly useful for patients. Arrangement Algorithms were considered for assessing their grouping execution as far as Accuracy, Precision, Sensitivity, and Specificity in characterizing the Wisconsin Breast cancer patient dataset. These measures can help the restorative experts in basic leadership all around effectively. We analyzed our data and experimented with the mean imputation method and k-Nearest Neighbor classification algorithm and obtained maximum accuracy. Comparative analysis of the classification of cancer diagnosis would provide further encouragement and efficient approaches for detecting the problems of cancer [34].

A paper that is the first study to evaluate breast cancer survival in Southern Iran and has used a wide range of explanatory factors, 44. The results demonstrate that survival is relatively poor and is associated with the diagnosis of late-stage disease. They hypothesize that this is due to a low level of awareness, lack of screening programs, ms, and subsequent late access to treatment. They got some results that look like that "The majority of patients were diagnosed with advanced tumor size. Five-year overall survival was 58% (95%CI; 53%–62%). Cox regression showed that family income (good vs poor: hazard ratio 0.46, 95%CI; 0.23–0.90) smoking (HR =1.40, 95%CI; 1.07–1.86), metastases to bone (HR = 2.25, 95%CI; 1.43–3.52) and lung (HR = 3.21, 95%CI;1.70–6.05), tumor size (2 cm vs 5 cm: HR = 2.07, 95%CI;1.39–3.09) and grade (poorly well-differentiated HR = 2.33, 95%CI; 1.52– 3.37), lymph node ratio (0 vs 1: HR = 15.31, 95%CI; 8.89–26.33) and number of involved nodes (1 vs >15: HR = 14.98, 95%CI; 8.83–25.33) were significantly related to survival [35].

By combining the machine learning methods K-NNs, Naive Bayes (NB), and reduced error pruning (REP) tree with the feature selection algorithm particle swarm optimization (PSO), Sakri et al. concentrated on improving the accuracy value. Their area of expertise includes the issue of breast cancer in Saudi Arabian women, which is one of the country's key issues, per their study. According to their reports, women over the age of 46 seem to be the main targets of this dangerous illness. The authors performed five phase-based data analysis approach on the WBCD dataset while maintaining this viewpoint. They published a comparison of classifiers that do not use

feature selection methods versus classifiers that do use feature selection methods. They have acquired 70%, 76.3%, and 66.3% accuracy for NB, RepTree, and K-NNs, respectively. They used Weka tool for their data analysis purpose. With PSO implemented, they have found four features that are best for this classification task. For NB, RepTree, and K-NN with PSO, they obtained 81.3%, 80%, and 75% accuracy values, respectively [36].

On the WBCD and another breast cancer dataset that was downloaded from the UCI library, Kapil and Rana adopted a modified decision tree technique they had proposed a weight-improved decision tree. They discovered that they have ranked each characteristic and retained the important features for this classification job using the Chisquare test. Their suggested method gained roughly 99% accuracy for the WBCD dataset and between 85% and 90% accuracy for the breast cancer dataset [37].

Azar et al. first described a decision tree-based technique for predicting breast cancer. The single decision tree (SDT), boosted decision tree (BDT), and decision tree forest is the modalities employed in this method (DTF). The choice is made after training the data set and testing the results. The results showed that the accuracy attained by SDT and BDT in the training phase was 97.07% and 98.83%, respectively. This clearly shows that BDT outperformed SDT. In the testing phase, decision tree forest achieved an accuracy of 97.51% versus SDT's 95.75%. Using ten-fold cross-validation, the dataset was trained. The authors illustrated a method for finding breast cancer. The experiments that have been done for detecting the disease are discussed here using local linear wavelet neural network (LLWNN), and recursive least square (RLS) to enhance the performance of the system. The LLWNN-RLS is providing the maximum values of average Correct Classification Rate (CCR) 0.897 and 0.972 for 2 and 3 predictors, respectively, with a few calculation times. It also provides the lowest value of minimum description length (MDL) and average squared classification error (ASCE) with much lesser time [38].
In a report, Md. Milon Islam and his team compared the performance of support vector machines, K-nearest neighbors, random forests, artificial neural networks, and logistic regression for the prediction of breast cancer. Each of the four machine learning techniques' fundamental characteristics and operations was demonstrated. In contrast to the lowest accuracy gained from RFs and LR, which is 95.7%, the maximum accuracy attained by ANNs is 98.57%. In the medical field, the diagnosing process is very time- and money-consuming. The system suggested that machine learning techniques might serve as a clinical aid for the detection of breast cancer and would be highly beneficial for newly qualified medical professionals or doctors in the event of a misdiagnosis. The developed model by ANNs is more consistent than any other technique stated, and it may be able to bring changes in the field of prediction of breast

cancer. From the study, we can conclude that machine learning techniques can detect the disease automatically with high accuracy [39].

In this study, they compared the performance of a conventional multiple CPH regression against three different ML methods (RSFs, SSVMs, and XGB) in a ranked survival prediction task using a dataset consisting of 36,658 Dutch non-metastatic breast cancer patients. Furthermore, they used SHAP values to open the models' black boxes and explain the difference in performance between a reference model (CPH) and the best-performing ML model (XGB). Their results showed that in the data at hand, ML-based approaches are capable of performing well as a conventional CPH model or, in the case of the XGB model, even better. However, this comes at the cost of an increase in complexity/opacity. ML explains ability techniques have been raised as a solution for this issue. They can help us generate an explicit knowledge representation of how the model makes its predictions. In their case, SHAP values showed that the key difference between CPH's and XGB's performance can be attributed, at least partially, to the latter's ability to capture data nonlinearities and interactions between features, which can have important contributions to the outputs. Moreover, it does so automatically and without any additional effort required by the researcher. Furthermore, SHAP values also allowed us to investigate the impact of specific features on the model predictions, which can be a complex task even for experts. This type of modeling framework could speed up the process of generating and testing new hypotheses on new (NCR) data, which could contribute to a rapid learning health system [40].

Chapter 4 Methodology

4.1 Data Collection and Analysis

4.1.1 Dataset-1:

The dataset chosen for this research is the Wisconsin Breast Cancer (Diagnostic) Data Set (WBCD). The dataset is publicly available on the reputed Machine Learning Repository which is UCI-Repository. WBCD was made by Dr. William H. Wolberg, a doctor at the University Of Wisconsin Hospital in Madison, Wisconsin, USA. Dr. Wolfsburg used Xcyt to analyze fluid samples taken from patients with solid breast masses [25]. Xcyt is an easy-to-use graphical computer program that is equipped to perform the investigation of cytological features based on digital scans. The dataset comprises o69 samples and 32 attributes of visually measured atomic features computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. FNA is a thin needle that is injected into the region of abnormal-appearing body fluid or tissues and collects a sample to make a diagnosis or predict diseases such as cancer. Among the 569 samples, the class distribution is 212 cancerous tumors (malignant) and other 357 non-cancerous tumors (benign). Ten features are computed from each one of the cells in the sample which are as follows [41]:

1. Radius (mean of distances from the center to points on the perimeter)
2. Texture (standard deviation of gray-scale values)
3. Perimeter
4. Area
5. Smoothness (local variation in radius lengths)
6. Compactness ($perimeter^2$ / area—1.0)
7. Concavity (severity of concave portions of the contour)
8. Concave points (number of concave portions of the contour)
9. Symmetry
10. Fractal dimension (“coastline approximation”—1)

The mean, standard error, and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, and field 23 is Worst Radius. All feature values are recorded with four significant digits missing attribute values: none

Class distribution: 357 benign, 212 malignant

Attributes in dataset-1**:**

i. **Diagnosis:** The diagnosis of breast tissues (M = malignant, B = benign)
ii. **Radius mean:** Mean of distances from the center to points on the perimeter
iii. **Texture means:** Standard deviation of gray-scale values
iv. **Perimeter mean:** Mean size of the core tumor
v. **Area mean**
vi. **Smoothness mean:** Mean of local variation in radius lengths
vii. **Compactness mean:** Mean of perimeter^2 / area - 1.0
viii. **Concavity mean:** Mean of the severity of concave portions of the contour
ix. **Concave points mean:** Mean for several concave portions of the contour
x. **Symmetry mean**
xi. **Fractal dimension mean:** Mean for "coastline approximation" - 1
xii. **Radius SE:** Standard error for the mean of distances from the center to points on the perimeter
xiii. **Texture SE:** Standard error for the standard deviation of gray-scale values
xiv. **Perimeter SE**
xv. **Area SE**
xvi. **Smoothness SE:** Standard error for local variation in radius lengths
xvii. **Compactness SE:** Standard error for perimeter^2 / area - 1.0
xviii. **Concavity SE:** Standard error for the severity of concave portions of the contour
xix. **Concave points se:** Standard error for several concave portions of the contour
xx. **Symmetry se:**
xxi. **Fractal dimension SE:** Standard error for "coastline approximation" - 1
xxii. **Radius worst:** "worst" or largest(mean of the three largest values) mean value for the mean of distances from the center to points on the perimeter
xxiii. **Texture worst:** "Worst" or largest mean value for the standard deviation of gray-scale values
xxiv. **Perimeter worst**
xxv. **Area worst**
xxvi. **Smoothness worst:** "Worst" or largest mean value for local variation in radius lengths
xxvii. **Compactness worst:** "Worst" or largest mean value for perimeter^2 / area - 1.0
xxviii. **Concavity worst:** "Worst" or largest mean value for the severity of concave portions of the contour
xxix. **Concave points worst:** "Worst" or largest mean value for several concave portions of the contour
xxx. **Symmetry worst**

xxxi. **Fractal dimension worst:** "Worst" or largest mean value for "coastline approximation" - 1

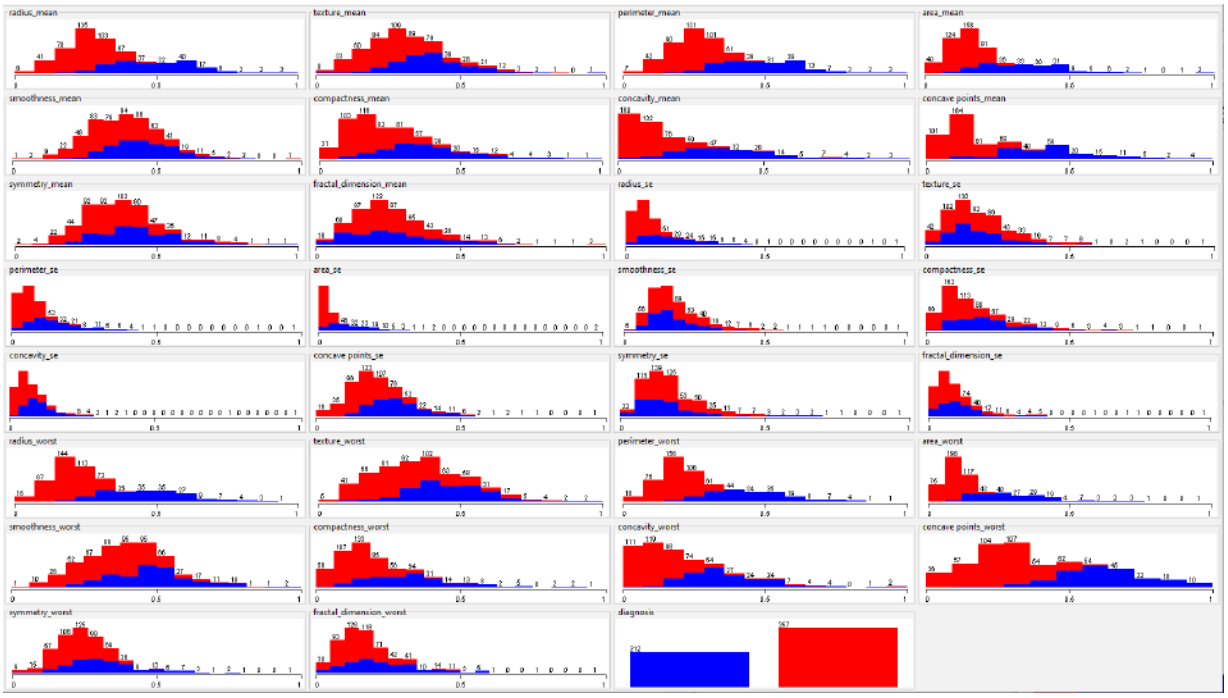


Figure 4: Attributes distribution (Dataset 1)

4.1.2 Dataset-2:

This database of breast cancer patients was acquired from the SEER Program of the NCI's November 2017 update, which offers details on population-based cancer statistics. The dataset included female patients who had been diagnosed between 2006 and 2010 with infiltrating ductal and lobular carcinoma breast cancer (SEER primary cites recode NOS histology codes 8522/3). In the end, 4024 patients were included after the exclusion of patients with uncertain tumor sizes, studied regional LNs, regional positive LNs, and patients whose survival months were less than 1 month [42].

Theoretical understanding of dataset-2

Stages of Cancer

In both staging systems, 7 key pieces of information are used [43] [44]:

The extent (size) of the tumor (T): How large is cancer? Has it grown into nearby areas?

The spread to nearby lymph nodes (N): Has cancer spread to nearby lymph nodes? If so, how many?

The spread (metastasis) to distant sites (M): Has cancer spread to distant organs such as the lungs or liver?

Estrogen Receptor (ER) status: Does cancer have the protein called an estrogen receptor?

Progesterone Receptor (PR) status: Does cancer have the protein called a progesterone receptor?

HER2 status: Does cancer make too much of a protein called HER2?

Grade of cancer (G): How much do the cancer cells look like normal cells?

TNM staging system

Tumor (T): How large is the primary tumor in the breast? What are its biomarkers?

Node (N): Has the tumor spread to the lymph nodes? If so, where, what size, and how many?

Metastasis (M): Has the cancer spread to other parts of the body?

Tumor (T):

T followed by a number from 0 to 4 describes the main (primary) tumor's size and if it has spread to the skin or the chest wall under the breast.

Higher T numbers mean a larger tumor and/or wider spread to tissues near the breast.

TX: Primary tumor cannot be assessed.

T0: No evidence of primary tumor.

Tis: Carcinoma in situ (DCIS, or Paget disease of the breast with no associated tumor mass)

T1 (includes T1a, T1b, and T1c): Tumor is 2 cm (3/4 of an inch) or less across

T2: Tumor is more than 2 cm but not more than 5 cm (2 inches) across.

T3: Tumor is more than 5 cm across.

T4 (includes T4a, T4b, T4c, and T4d): Tumor of any size growing into the chest wall or skin. This includes inflammatory breast cancer.

Node (N):

The "N" in the TNM staging system stands for lymph nodes. These small, bean-shaped organs help fight infection.

Node (N) describes whether the cancer has spread to the lymph nodes.

Lymph nodes near where the cancer started are called regional lymph nodes.

Lymph nodes in other parts of the body are called distant lymph nodes.

If the lymph nodes feel enlarged, it's likely the breast cancer has spread.

Lymph nodes located under the arm called the axillary lymph nodes

Lymph nodes located under the breastbone called the internal mammary lymph nodes

If the lymph nodes feel enlarged, it's likely the breast cancer has spread.

N followed by a number from 0 to 3 indicates whether the cancer has spread to lymph nodes near the breast and, if so, how many lymph nodes are involved.

N1: Cancer has spread to 1 to 3 axillary (underarm) lymph node(s), and/or cancer is found in internal mammary lymph nodes (those near the breast bone) on sentinel lymph node biopsy.

N2: Cancer has spread to 4 to 9 lymph nodes under the arm, or cancer has enlarged the internal mammary lymph nodes

N3: The cancer has spread to 10 or more axillary lymph nodes, or it has spread to the lymph nodes located under the clavicle, or collarbone. It may have also spread to the internal mammary lymph nodes.

Cancer that has spread to the lymph nodes above the clavicle, called the supraclavicular lymph nodes, is also described as N3.

Metastasis (M)

M followed by a 0 or 1 indicates whether the cancer has spread to distant organs – for example, the lungs, liver, or bones.

M0: No distant spread is found on x-rays (or other imaging tests) or by physical exam.

M0(i+): Small numbers of cancer cells are found in blood or bone marrow (found only by special tests), or tiny areas of cancer spread (no larger than 0.2 mm) are found in lymph nodes away from the underarm, collarbone, or internal mammary areas.

M1: Cancer has spread to distant organs (most often to the bones, lungs, brain, or liver) as seen on imaging tests or by physical exam, and/or a biopsy of one of these areas proves cancer has spread and is larger than 0.2mm.

Stage groups for breast cancer

- **Stage 0** involves only a small cluster of cancer cells in the duct or lobule.
- **Stage 1** is a tumor smaller than 2 cm.
- **Stage 2** is a tumor up to 5 cm that has not spread to the axillary lymph nodes.
- **Stage 3** is a tumor of any size that may have spread to the axillary lymph nodes.
- **Stage 4** is a tumor of any size that has metastasized and has gone to other tissues besides the breast and lymph nodes.

Doctors assign the stage of the cancer by combining the T, N, and M classifications (see above), the tumor grade, and the results of ER/PR and HER2 testing.

Stage 0: Stage zero (0) describes a disease that is only in the ducts of the breast tissue and has not spread to the surrounding tissue of the breast. It is also called non-invasive or in situ cancer (Tis, N0, M0).

Stage IA: The tumor is small, invasive, and has not spread to the lymph nodes (T1, N0, M0).

Stage IB: Cancer has spread to the lymph nodes and the cancer in the lymph node is larger than 0.2 mm but less than 2 mm in size. There is either no evidence of a tumor in the breast or the tumor in the breast is 20 mm or smaller (T0 or T1, N1mi, M0).

Stage IIA: Any 1 of these conditions:

There is no evidence of a tumor in the breast, but the cancer has spread to 1 to 3 axillary lymph nodes. It has not spread to distant parts of the body (T0, N1, M0).

The tumor is 20 mm or smaller and has spread to 1 to 3 axillary lymph nodes (T1, N1, M0).

The tumor is larger than 20 mm but not larger than 50 mm and has not spread to the axillary lymph nodes (T2, N0, M0).

Stage IIB: Either of these conditions:

The tumor is larger than 20 mm but not larger than 50 mm and has spread to 1 to 3 axillary lymph nodes (T2, N1, M0).

The tumor is larger than 50 mm but has not spread to the axillary lymph nodes (T3, N0, M0).

Stage IIIA: The tumor of any size has spread to 4 to 9 axillary lymph or internal mammary lymph nodes. It has not spread to other parts of the body (T0, T1, T2, or T3; N2; M0). Stage IIIA may also be a tumor larger than 50 mm that has spread to 1 to 3 axillary lymph nodes (T3, N1, M0).

Stage IIIB: The tumor has spread to the chest wall or caused swelling or ulceration of the breast, or it is diagnosed as inflammatory breast cancer. It may or may not have spread to up to 9 axillary or internal mammary lymph nodes. It has not spread to other parts of the body (T4; N0, N1, or N2; M0).

Stage IIIC: A tumor of any size that has spread to 10 or more axillary lymph nodes, the internal mammary lymph nodes, and/or the lymph nodes under the collarbone. It has not spread to other parts of the body (any T, N3, M0).

Stage IV (metastatic): The tumor can be any size and has spread to other organs, such as the bones, lungs, brain, liver, distant lymph nodes, or chest wall (any T, any N, M1).

Grade and Differentiation

The grade describes the appearance of the cancerous cells.

What does grade mean in a cancer diagnosis?

Tumor grade describes how normal or abnormal cancer cells look under a microscope. The more normal the cells look, the less aggressive the cancer and the more slowly it grows and spreads. On the other hand, the more abnormal the cells look the more aggressive the cancer and the faster it is likely to grow and spread [45] [46]

Grade 1: Tumor cells and tissue looks most like healthy cells and tissue. These are called well-differentiated tumors and are considered low grade.

Grade 2: The cells and tissue are somewhat abnormal and are called moderately differentiated. These are intermediate grade tumors.

Grade 3: Cancer cells and tissue look very abnormal. These cancers are considered poorly differentiated since they no longer have an architectural structure or pattern. Grade 3 tumors are considered high grade.

Grade 4: These undifferentiated cancers have abnormal-looking-looking cells. These are the highest grade and typically grow and spread faster than lower-grade tumors.

What does it mean if your breast cancer is estrogen and progesterone positive?

Hormone Receptor-Positive Breast Cancer

About 80% of all breast cancers are “ER-positive.” That means the cancer cells grow in response to the hormone estrogen. About 65% of these are also “PR-positive.” They grow in response to another hormone, progesterone [47] [48]

How do estrogen and progesterone affect breast cancer?

Breast cancer cells taken out during a biopsy or surgery will be tested to see if they have certain proteins that are estrogen or progesterone receptors. When the hormones estrogen and progesterone attach to these receptors, they stimulate the cancer to grow [49].

Studies have also shown that a woman’s risk of breast cancer is related to the estrogen and progesterone made by her ovaries (known as endogenous estrogen and progesterone). Being exposed for a long time and/or to high levels of these hormones has been linked to an increased risk of breast cancer.

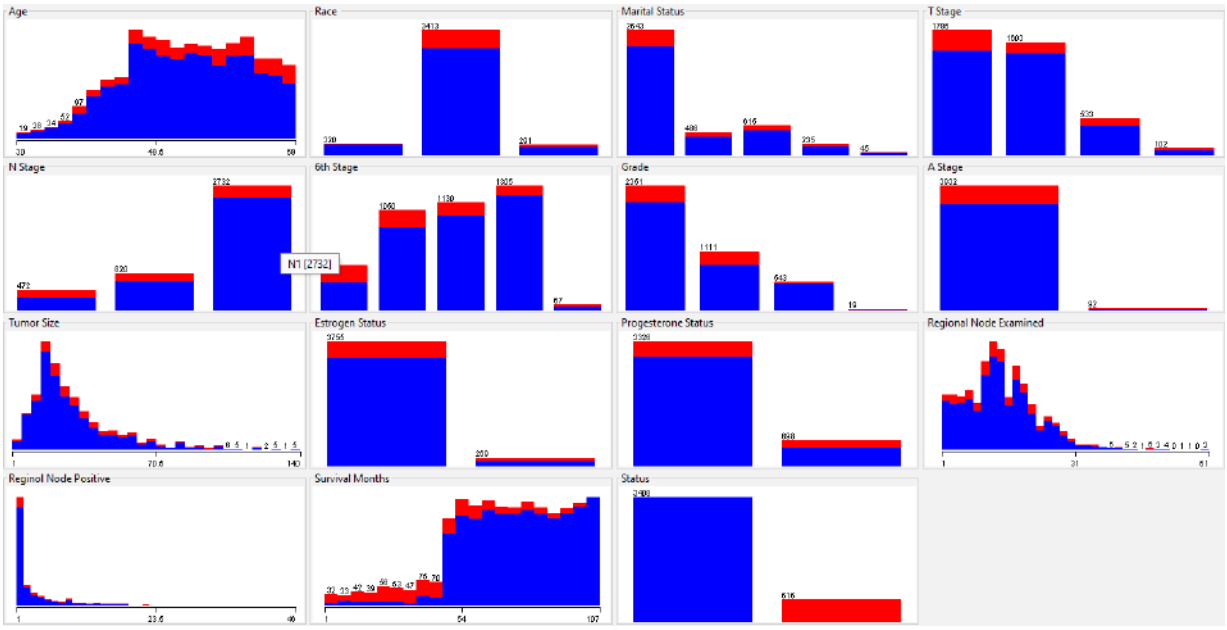


Figure 5: Attribute distribution (Dataset 2)

Attributes Analysis:

Age:

From the data table, we can see the range of age of the population and also the number. The range is divided into 4 classes with a range of 10 years from 30 - 69. The percentage of cancer affected is highest in the 50-59 years range [50]

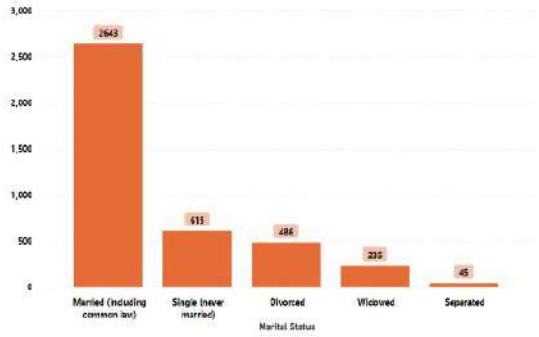
Here we can see the average age of cancer patients is 53.97.

Race:

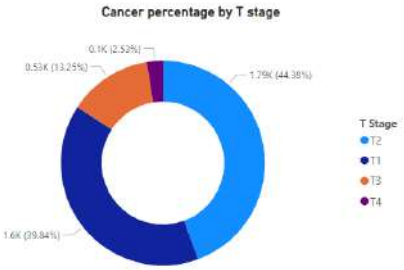
The data is divided into three races: White, Black & Other.
In this data, the number of white patients is higher than any other aches [51].

Marital Status:

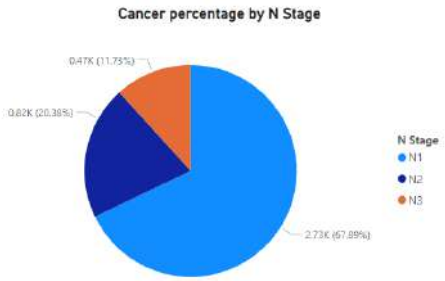
This data shows the marital status of the patients. Here we can see the percentage of affected people is higher in married women, which is about (65.7%).



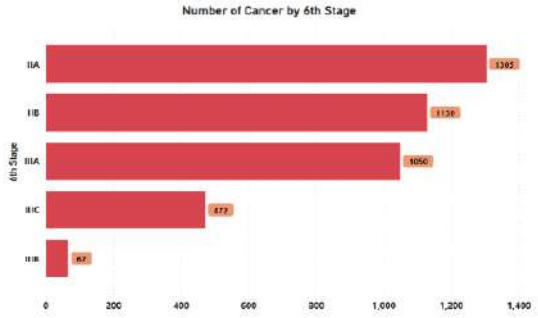
T Stage:
T followed by a number from 0 to 4 describes the main (primary) tumor's size and if it has spread to the skin or the chest wall under the breast.
The percentage of T stage is T2 (44.4%), T1(39.74%), T3 (13.2%), and T4 (2.48%) respectively.



N Stage: N followed by a number from 0 to 3 indicates whether the cancer has spread to lymph nodes near the breast and, if so, how many lymph nodes are involved.
N-stage percentages are N1 (67.9%), N2 (20.4%), and N3 (11.7%) respectively.



6th Stage: Doctors assign the stage of the cancer by combining the T, N, and M classifications.
Our dataset provides us with five different 6th stage information: IIA, IIB, IIIA, IIIB, IIIC.
The percentage of 6th stages are IIA (32.4%), IIB (28%), IIIA (26%), IIIB (1.67%), IIIC (11.73%).



Grade: The grade describes the appearance of the cancerous cells. Which is referred to how normal or abnormal cancer cells look under a microscope.

The percentage of grading among the data is as follows,

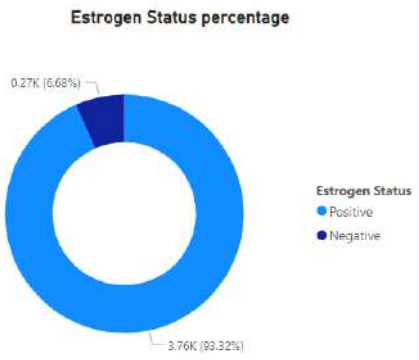
- Grade - 13.5%,
- Grade II - 58.4%
- Grade III - 27.6%
- Grade IV - 0.47%.

A Stage:
It refers to the spreading area of the cancer cells. That means if the cancer cell has spread near or far from the primary cancer cell
Regional – Cell around the area
Distant – Cell far from the main area

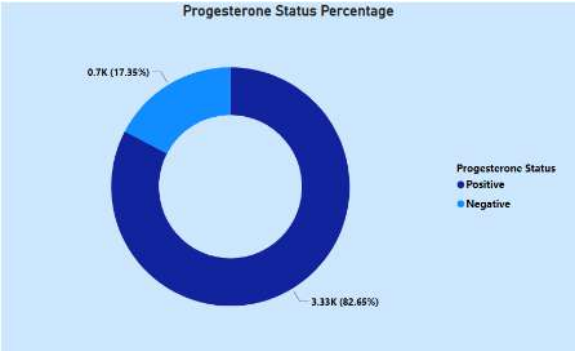
Tumor Size:

It refers to the spread of cancer cells. That is, how big is the tumor?
Here the percentages are
[<36 mm] - 73%
[36 mm - 70 mm] – 21%
[71 mm - 105 mm] – 4.6%
[>105 mm] – 0.89%

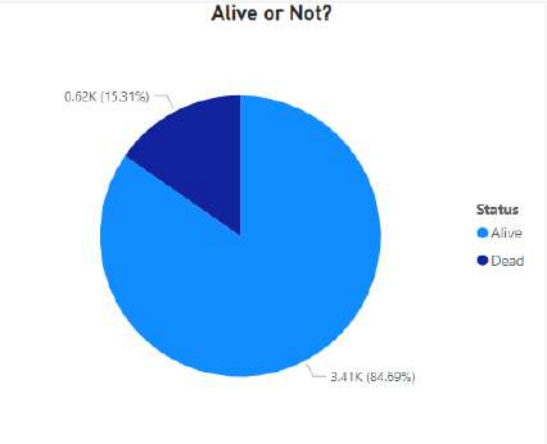
Estrogen Status:
That means the cancer cells grow in response to the hormone estrogen.
Among the patients, Estrogen Positive is 93.3% & Negative is 6.68%.



Progesterone Status:
That means the cancer cells grow in response to the hormone progesterone.
In this set of data Progesterone Positive is 82.7% & Negative is 17.3%.



Living Status:
This Data gives us information on whether the patients are alive or dead.
From the data, we can see that 84.7% of the patients are alive and 15.3% of the patients are dead [52].



Death Rate Analysis:

Age: (Death Rate)

30-39: Death Rate (47/230) = 20.43%

40-49: Death Rate (136/1124) = 12.09%

50-59: Death Rate (189/1390) = 13.59%

60-69: Death Rate (244/1280) = 19.06%

- **Young and Old Women have higher death rates than middle-aged women.**

Race: (Death Rate)

Black: Death Rate (73/291) = 25%

White: Death Rate (510/3413) = 14.94%

Other: Death Rate (33/320) = 10.31%

- **Black Women survive less than White or Other people.**

Marital Status: (Death Rate)

Divorced: Death Rate (90/486) = 18.51%

Married: Death Rate (358/2643) = 13.54%

Separated: Death Rate (15/45) = 33.33%

Single: (104/615) = 16.91%

Widowed: (49/235) = 20.85%

- **Separated Women survive less in breast cancer**

T Stage: (Death Rate)

T1: (157/1603) = 9.79%

T2: (303/1786) = 16.96%

T3: (116/533) = 21.76%

T4: (40/102) = 39.21%

- **T4 stage is the most dangerous T stage, the death rate is 39.21% in this stage**

N Stage (Death Rate)

N1: (270/2732) = 9.88%

N2: (165/820) = 20.12%

N3: (181/472) = 38.34%

- **The Death rate in N3 stage is 38.34%**

6th Stage (Death Rate)

IIA: (96/1305) = 7.35%

IIB: (135/1130) = 11.94%

IIIA: (184/1050) = 17.52%

IIIB: (20/67) = 29.85%

IIIC: (181/472) = 38.34%

- **The death rate in the IIIC stage is 38.34%**

So, the earlier stage we can detect the disease, the better chances are there to survive

Grade (Death Rate)

Grade I: (39/543) = 7.18%

Grade II: (305/2351) = 12.97%

Grade III: (263/1111) = 23.67%

Grade IV (Undifferentiated): (9/19) = 47.37%

- **Undifferentiating increases the risk of death (47.37%)**

A Stage (Death Rate)

Distant: (35/92) = 38.04%

Regional: (581/3932) = 14.77%

- **Distant Stage has a higher death rate (38.04%)**

Tumor Size (Death Rate)

<36: (383/2956) = 12.95%

36-70: (172/846) = 20.33%

71-105: (51/186) = 27.41%

>105: (10/36) = 27.78%

- **The bigger the tumor the higher the Death risk**

Estrogen Status (Death Rate)

Positive: (508/3755) = 13.52%

Negative: (108/269) = 40.14%

- **Estrogen being positive is important, otherwise increases the death rate (40.14%)**

Progesterone Status (Death Rate)

Positive: (412/3326) = 12.38%

Negative: (204/698) = 29.22%

- **Progesterone being negative is also risky (29.22% death rate)**

Regional Node Examined (Death Rate)

<16: (346/2389) = 14.48%

16-30: (246/1500) = 16.4%

31-45: (19/119) = 15.96%

>45: (5/16) = 31.25%

Regional Node Positive (Death Rate)

<13: (493/3730) = 13.21%

13-22: (93/232) = 40.08%

23-33: (27/55) = 49.09%

>33: (3/7) = 42.85%

- **The more nodes affected the higher risk of death.**

Survival Months (Death Rate)

<28: (159/215) = 73.95%

29-54: (252/677) = 37.22%

55-80: (139/1577) = 8.81%

>80: (66/1555) = 4.24%

The more a patient survives the higher chances of being cured. Surviving 28 months could be a milestone as 73.95% of patients died before 28 months.

4.3 Data Pre-processing

Dataset-1

We've preprocessed the dataset in the following way:

- **Remove Unwanted Feature:** 'Id' feature is unnecessary, so we shall remove this.
- **Normalization:** After that, we normalized the dataset using the min-max normalization technique.

Dataset-2

The dataset was clean and processed. We've just followed the below processing:

- **Discretization**: We've converted the numeric values into Nominal by discretizing the attribute into 4 groups. Age, Tumor Size, Regional Node Examined and Regional Node Positive has been discretized within the following range:

Attribute	Discretized Group
Age	30-39, 40-49, 50-59, 60-69
Tumor Size	<36, 36-70, 71-105, >105
Regional Node Examined	<16, 16-30, 31-45, >45
Regional Node Positive	<13, 13-22, 23-33, >33

- **Imbalance Handling:** The class distribution shows some imbalance issues. We've tried to handle these imbalance issues in 3 ways by using Class Balancer, SMOTe, and under-sampling by Spread Subsample technique.

4.4 Feature Selection:

Dataset-1

We've 31 features in total. We want to apply some feature selection techniques to select the important features only. We've applied Info Gain Attribute Evaluation, Gain Ratio Attribute Evaluation with Ranker method with threshold 0 an, and CFS Attribute Evaluation with the best first method. We've got finally, 4 datasets Raw Data, Info Gain Data (IGD), CFS Selection Data (CFS), and Gain Ratio Data (GRD).

CFS Subset Result:

```
=== Attribute Selection on all input data ===

Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 336
        Merit of best subset found:    0.667

Attribute Subset Evaluator (supervised, Class (nominal): 31 diagnosis):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 2,7,8,14,19,21,23,24,25,27,28 : 11
                     texture_mean
                     concavity_mean
                     concave points_mean
                     area_se
                     symmetry_se
                     radius_worst
                     perimeter_worst
                     area_worst
                     smoothness_worst
                     concavity_worst
                     concave points_worst
```

Info Gain Result:

```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 31 diagnosis):
        Information Gain Ranking Filter

Ranked attributes:
 0.685    23 perimeter_worst
 0.6686   24 area_worst
 0.6665   21 radius_worst
 0.6478   28 concave points_worst
 0.6347    8 concave points_mean
 0.5623    3 perimeter_mean
 0.5479    4 area_mean
 0.541     1 radius_mean
 0.5171    7 concavity_mean
 0.517    14 area_se
 0.4735   27 concavity_worst
 0.3679   11 radius_se
 0.3663   13 perimeter_se
 0.3204   26 compactness_worst
 0.304     6 compactness_mean
 0.2225   17 concavity_se
 0.197    18 concave points_se
 0.1881   22 texture_worst
 0.1593    2 texture_mean
 0.1492   29 symmetry_worst
 0.1303   16 compactness_se
 0.1235   25 smoothness_worst
 0.0988    9 symmetry_mean
 0.0971    5 smoothness_mean
 0.0747   30 fractal_dimension_worst
 0.0346   20 fractal_dimension_se
 0.0228   19 symmetry_se
 0        10 fractal_dimension_mean
 0        12 texture_se
 0        15 smoothness_se

Selected attributes: 23,24,21,28,8,3,4,1,7,14,27,11,13,26,6,17,18,22,2,29,16,25,9,5,30,20,19,10,12,15 : 30
```

Gain Ratio Result:

```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 31 diagnosis):
        Gain Ratio feature evaluator

Ranked attributes:
 0.4443   23 perimeter_worst
 0.3872   21 radius_worst
 0.3851   24 area_worst
 0.3844   28 concave points_worst
 0.3288    8 concave points_mean
 0.3195    7 concavity_mean
 0.3127   27 concavity_worst
 0.3123    3 perimeter_mean
 0.3061    4 area_mean
 0.2968    1 radius_mean
 0.266    14 area_se
 0.2156    6 compactness_mean
 0.215    11 radius_se
 0.194    13 perimeter_se
 0.1681   26 compactness_worst
 0.1604   17 concavity_se
 0.1595    2 texture_mean
 0.145    19 symmetry_se
 0.1281   18 concave points_se
 0.1256   25 smoothness_worst
 0.1201   22 texture_worst
 0.1072   29 symmetry_worst
 0.1053    5 smoothness_mean
 0.1024   16 compactness_se
 0.0956   30 fractal_dimension_worst
 0.0682    9 symmetry_mean
 0.0346   20 fractal_dimension_se
 0        12 texture_se
 0        10 fractal_dimension_mean
 0        15 smoothness_se

Selected attributes: 23,21,24,28,8,7,27,3,4,1,14,6,11,13,26,17,2,19,18,25,22,29,5,16,30,9,20,12,10,15 : 30
```

Applying each feature selection method shall help us in performing the algorithms over each of the different datasets to find the best feature selection method in this case.

Dataset-2

We have 15 features which is not much in case but yet we want to select only the important ones. So, we tried the Info Gain Attribute Evaluation technique but it seems that all the features are important in our case and it might not be good to remove any of these attributes, the following figure shows the results obtained from info gain:

```
Attribute Evaluator (supervised, Class (nominal): 15 Status):
        Information Gain Ranking Filter

Ranked attributes:
 0.1506   14 Survival Months
 0.04883   6 6th Stage
 0.04734   5 N Stage
 0.02518  13 Reginol Node Positive
 0.0213    7 Grade
 0.01848  11 Progesterone Status
 0.01708  10 Estrogen Status
 0.01553   4 T Stage
 0.007     9 Tumor Size
 0.00666   3 Marital Status
 0.00651   1 Age
 0.00581   2 Race
 0.00502   8 A Stage
 0.00204  12 Regional Node Examined

Selected attributes: 14,6,5,13,7,11,10,4,9,3,1,2,8,12 : 14
```

4.5 Model Building

We have set a total of 6 algorithms for our model building for both datasets.

The algorithms are:

- Decision Tree (J48)
- K- Nearest Neighbor (KNN)
- Logistic Regression
- Naïve Bayes
- Random Forest
- Support Vector Machine

After building the models we evaluated them with two different evaluation methods.

4.6 Evaluation Method

Cross Validation (For Dataset-1):

Cross-validation is a technique for evaluating ML models by training several ML models on subsets of the available input data and evaluating them on the complementary subset of the data [53].

Cross-validation is a statistical method used to estimate the performance (or accuracy) of machine learning models. It is used to protect against overfitting in a predictive model, particularly in a case where the amount of data may be limited. In cross-validation, you make a fixed number of folds (or partitions) of the data, run the analysis on each fold, and then average the overall error estimate.

The purpose of cross-validation is to test the ability of a machine-learning model to predict new data. It is also used to flag problems like overfitting or selection bias and gives insights on how the model will generalize to an independent dataset.

There are 4 Types of Cross-Validation in Machine Learning:

- Holdout Method
- K-Fold Cross-Validation
- Stratified K-Fold Cross-Validation
- Leave-P-Out Cross-Validation

We used the K-fold cross-validation as our evaluation method. Where K=10.

K fold cross validation:

K-fold cross-validation is one way to improve the holdout method. This method guarantees that the score of our model does not depend on the way we picked the train and test set. The data set is divided into k=10 number of subsets and the holdout method is repeated k=10 number of times. Let us go through this in steps:

1. We randomly split our entire dataset into k=10 number of folds (subsets)
2. For each fold in our dataset, we built a model on k – 1 (9 in our case) folds of the dataset. Then, test the model to check the effectiveness for the kth (10^{th}) fold
3. We repeated this until each of the k-folds has served as the test set
4. The average of your k recorded accuracy is called the cross-validation accuracy and will serve as your performance metric for the model.

Because it ensures that every observation from the original dataset has the chance of appearing in the training and test set, this method generally results in a less biased model compared to other methods. It is one of the best approaches if we have limited input data.

The disadvantage of this method is that the training algorithm has to be rerun from scratch k times, which means it takes k times as much computation to make an evaluation.

Percentage split (For Dataset-2)

Train-Test Split Evaluation

The train-test split is a technique for evaluating the performance of a machine-learning algorithm [54].

It can be used for classification or regression problems and can be used for any supervised learning algorithm.

The procedure involves taking a dataset and dividing it into two subsets. The first subset is used to fit the model and is referred to as the training dataset. The second subset is not used to train the model; instead, the input element of the dataset is provided to the model, then predictions are made and compared to the expected values. This second dataset is referred to as the test dataset.

- **Train Dataset**: Used to fit the machine learning model.
- **Test Dataset**: Used to evaluate the fit machine learning model.

The objective is to estimate the performance of the machine learning model on new data: data not used to train the model.

The train-test procedure is appropriate when there is a sufficiently large dataset available.

Nevertheless, common split percentages include:

- Train: 80%, Test: 20%
- Train: 70%, Test: 30%
- Train: 67%, Test: 33%
- Train: 50%, Test: 50%

Very often, the proportion chosen is 70% for the training set and 30% for the test. The idea is that more training data is a good thing because it makes the classification model better whilst more test data makes the error estimate more accurate.

We have chosen the Train: 70%, Test: 30% splitting evaluation method for our work because we've handled the imbalance issues and we want to have a generalized outcome.

4.7 Performance Metrics

The performance of machine learning techniques is measured concerning a few performance metrics [55].

Confusion Metrics

Evaluation of the performance of a classification model is based on the counts of test records correctly and incorrectly predicted by the model. The confusion matrix provides a more insightful picture which is not only the performance of a predictive model, but also which classes are being predicted correctly and incorrectly, and what type of errors are being made. To illustrate, we can see how the 4 classification metrics are calculated (TP, FP, FN, TN), and our predicted value compared to the actual value in a confusion matrix is presented in the below confusion matrix table [55].

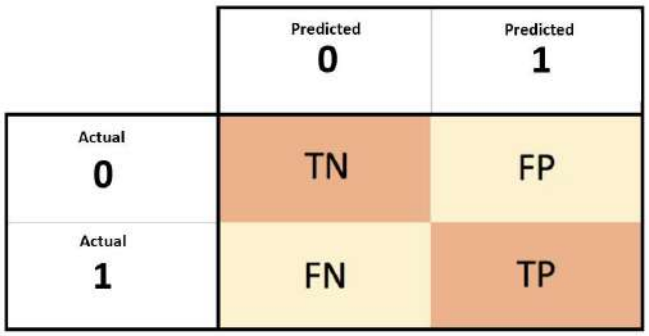


Figure 6: Possible Classification Outcomes: TP, FP, FN, TN.

- True Positive (TP) — a class is predicted true and is true in reality
- True Negative (TN) — a class is predicted false and is false in reality
- False Positive (FP) — a class is predicted true but is false in reality
- False Negative (FN) — a class is predicted false but is true in reality

In our study, the following parameters are used extensively to evaluate some terms by their corresponding formula to measure the performance of our study. There are a lot of parameters like these which describe some relationships that can help to measure the performance of a system. The comparative study's performance is evaluated by the following formulas:

Accuracy

The accuracy metric is one of the simplest Classification metrics to implement, and it can be determined as the number of correct predictions to the total number of predictions [56].

It can be formulated as:

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions}$$

Precision

The precision metric is used to overcome the limitation of Accuracy. The precision determines the proportion of positive prediction that was correct. It can be calculated as the True Positive or predictions that are true to the total positive predictions (True Positive and False Positive) [56].

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

Recall

It is also similar to the Precision metric; however, it aims to calculate the proportion of actual positive that was identified incorrectly. It can be calculated as a True Positive or a prediction that is true to the total number of positives, either correctly predicted as positive or incorrectly predicted as negative (true Positive and false negative).

The formula for calculating Recall is given below:

$$Recall = \frac{TP}{TP\ +\ FN}$$

F-Scores (F-Measure)

F-score or F1 Score is a metric to evaluate a binary classification model based on predictions that are made for the positive class. It is calculated with the help of Precision and Recall. It is a type of single score that represents both Precision and Recall. So, the F1 Score can be calculated as the harmonic mean of both precision and Recall, assigning equal weight to each of them [56].

The formula for calculating the F1 score is given below:

$$F1 - Score = 2 * \frac{precision * recall}{precision + recall}$$

4.8 Classification Algorithm Models:

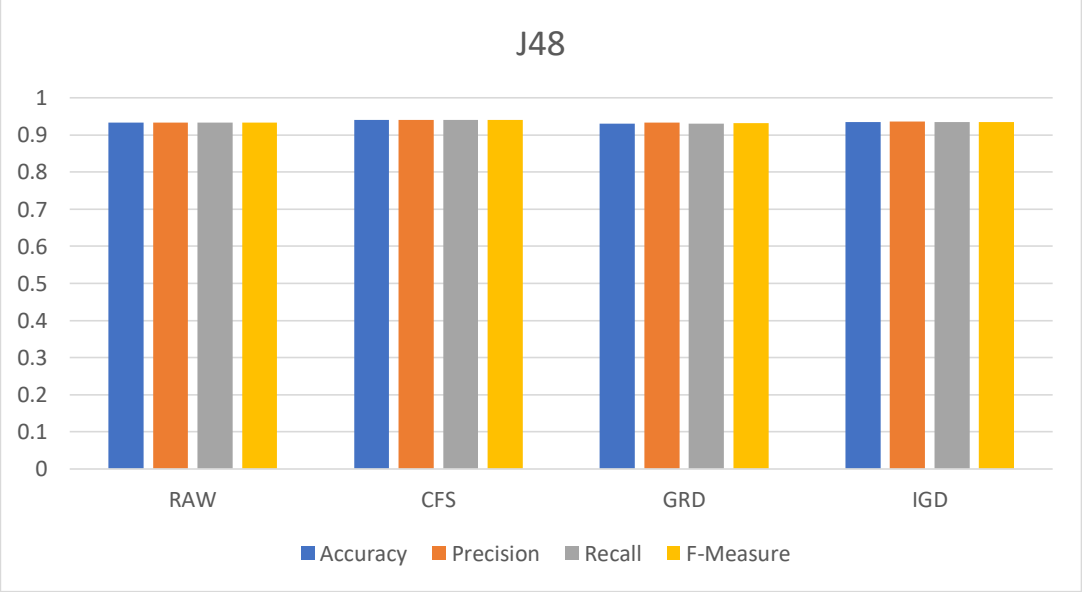
6 different algorithms are implemented on both of the datasets to predict the death rate and the stages of breast cancer.

Decision Tree(J48) Model

Dataset 1:

Table 1: Class-wise Results for J48

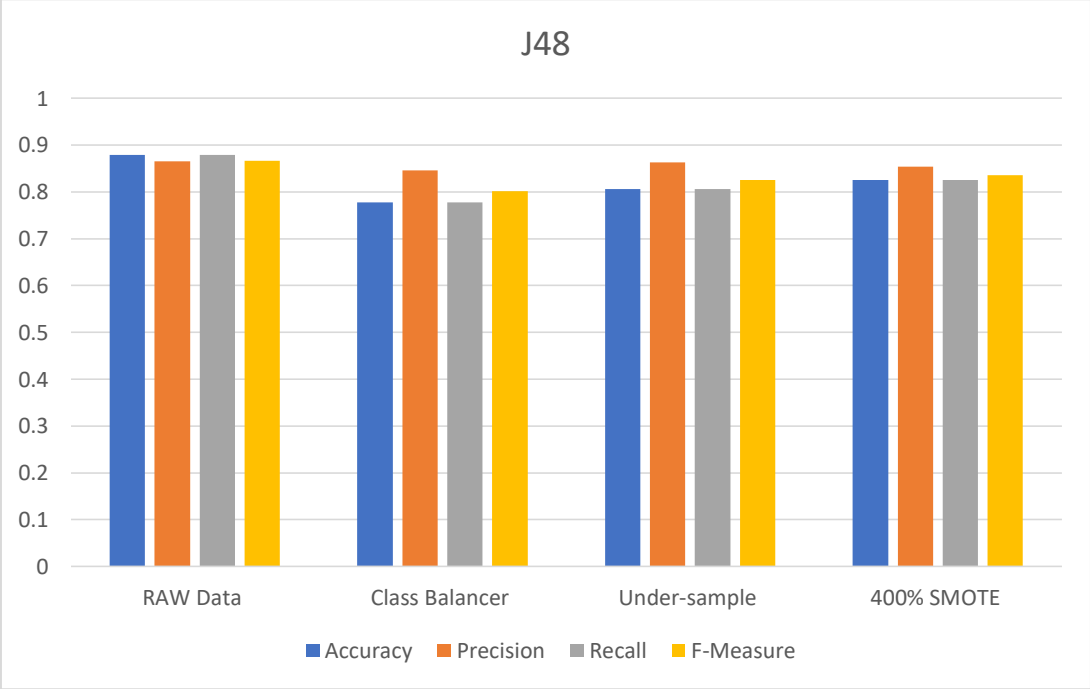
J48					
	Class	Accuracy	Precision	Recall	F-Measure
RAW	M	0.929	0.895	0.929	0.912
	B	0.936	0.957	0.936	0.946
	Avg.	0.933	0.934	0.933	0.933
CFS	M	0.939	0.905	0.939	0.921
	B	0.941	0.963	0.941	0.952
	Avg.	0.94	0.941	0.940	0.940
GRD	M	0.934	0.888	0.934	0.910
	B	0.930	0.96	0.930	0.945
	Avg.	0.931	0.933	0.931	0.932
IGD	M	0.934	0.896	0.934	0.915
	B	0.936	0.96	0.936	0.948
	Avg.	0.935	0.936	0.935	0.935



Dataset 2:

Table 2: Class-wise Results for J48

J48					
	Class	Accuracy	Precision	Recall	F-Measure
RAW Data	Alive	0.962	0.902	0.962	0.931
	Dead	0.409	0.655	0.409	0.503
	Avg.	0.879	0.865	0.879	0.867
Class Balancer	Alive	0.798	0.931	0.798	0.860
	Dead	0.663	0.367	0.663	0.472
	Avg.	0.778	0.846	0.778	0.802
Under-sample	Alive	0.823	0.942	0.823	0.878
	Dead	0.713	0.415	0.713	0.524
	Avg.	0.806	0.863	0.806	0.825
400% SMOTE	Alive	0.861	0.928	0.861	0.893
	Dead	0.619	0.439	0.619	0.514
	Avg.	0.825	0.854	0.825	0.836

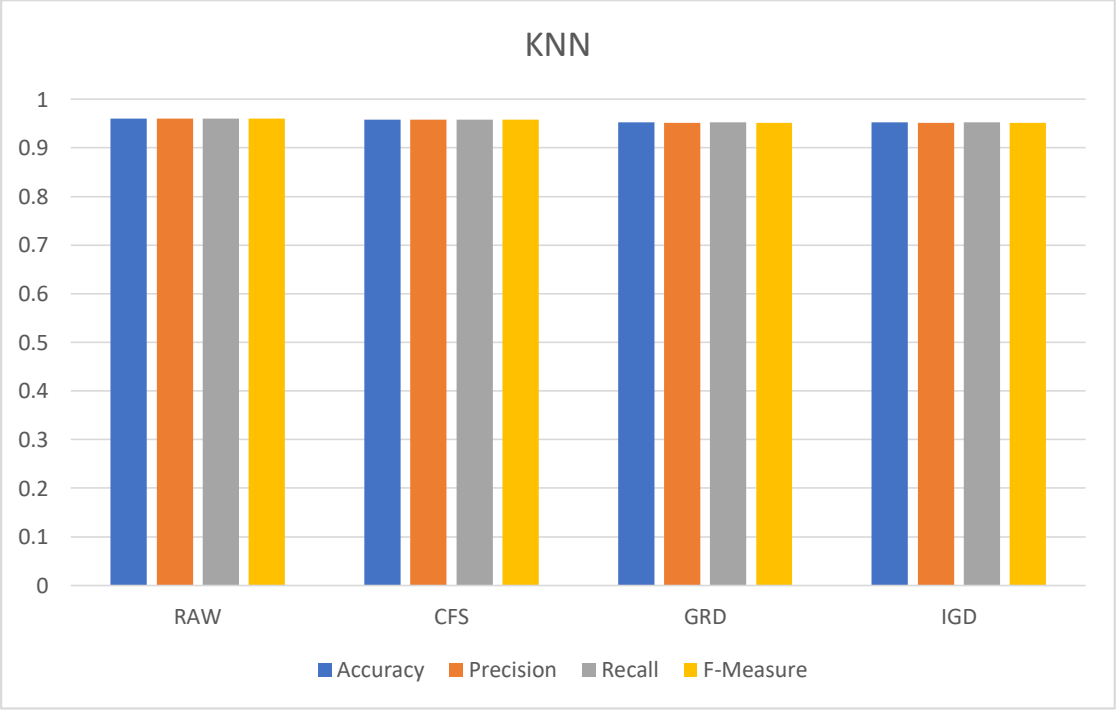


K-Nearest Neighbor (KNN) Model

Dataset 1:

Table 3: Class-wise Results for KNN

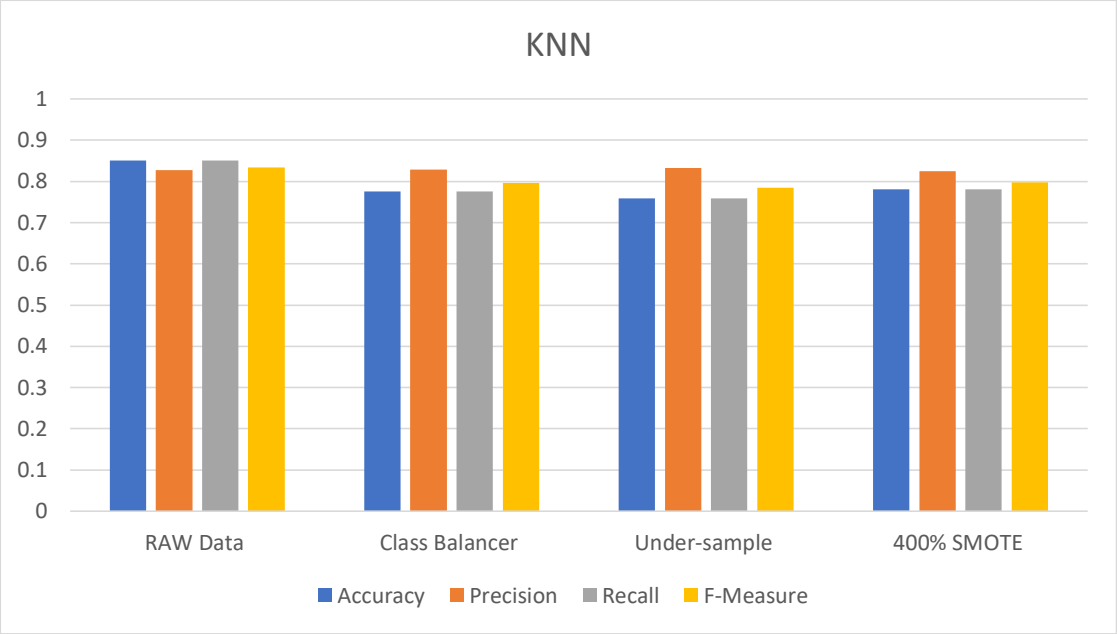
KNN					
	Class	Accuracy	Precision	Recall	F-Measure
RAW	M	0.943	0.948	0.943	0.946
	B	0.969	0.966	0.969	0.968
	Avg.	0.96	0.96	0.96	0.96
CFS	M	0.939	0.948	0.939	0.943
	B	0.969	0.964	0.969	0.966
	Avg.	0.958	0.958	0.958	0.958
GRD	M	0.929	0.943	0.929	0.936
	B	0.966	0.958	0.966	0.962
	Avg.	0.953	0.952	0.953	0.952
IGD	M	0.929	0.943	0.929	0.936
	B	0.966	0.958	0.966	0.962
	Avg.	0.953	0.952	0.953	0.952



Dataset 2:

Table 4: Class-wise Results for KNN

KNN					
	Class	Accuracy	Precision	Recall	F-Measure
RAW Data	Alive	0.949	0.884	0.949	0.915
	Dead	0.293	0.505	0.293	0.371
	Avg.	0.851	0.827	0.851	0.834
Class Balancer	Alive	0.814	0.913	0.814	0.861
	Dead	0.558	0.346	0.558	0.427
	Avg.	0.776	0.828	0.776	0.796
Under-sample	Alive	0.785	0.920	0.785	0.847
	Dead	0.613	0.334	0.613	0.433
	Avg.	0.759	0.832	0.759	0.785
400% SMOTE	Alive	0.825	0.909	0.825	0.865
	Dead	0.530	0.348	0.530	0.420
	Avg.	0.781	0.825	0.781	0.798

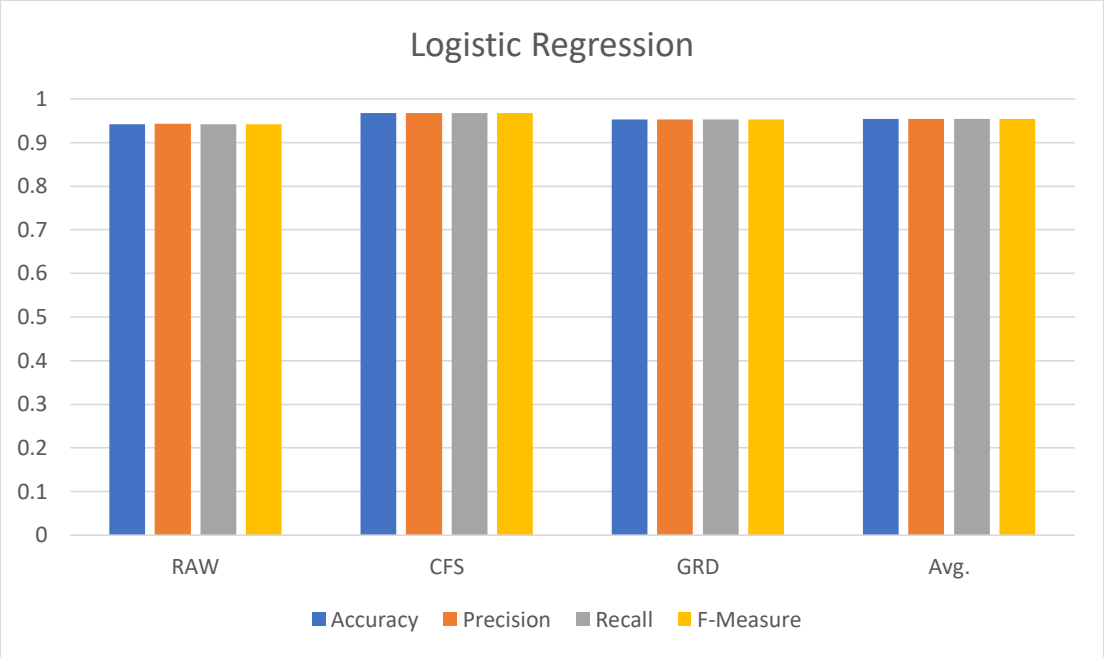


Logistic Regression Model:

Dataset 1:

Table 5: Class-wise Results for LR

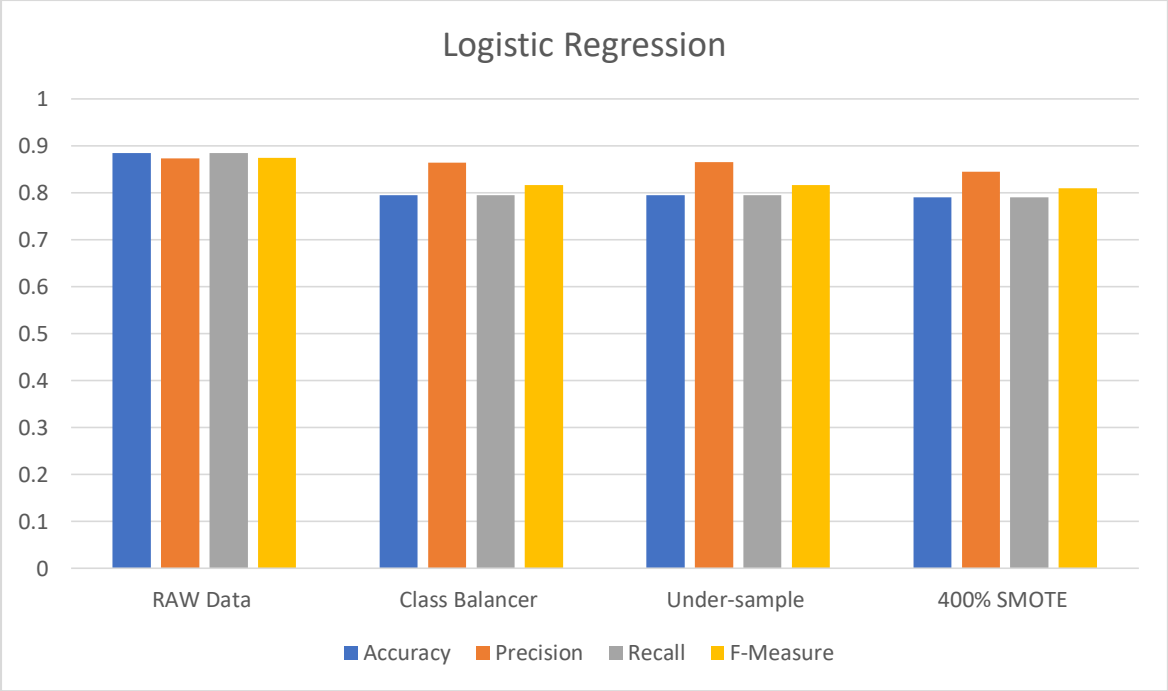
Logistic Regression					
	Class	Accuracy	Precision	Recall	F-Measure
RAW	M	0.939	0.909	0.939	0.923
	B	0.944	0.963	0.944	0.953
	Avg.	0.942	0.943	0.942	0.942
CFS	M	0.958	0.958	0.958	0.958
	B	0.975	0.975	0.975	0.975
	Avg.	0.968	0.968	0.968	0.968
GRD	M	0.934	0.938	0.934	0.936
	B	0.964	0.961	0.964	0.962
	Avg.	0.953	0.953	0.953	0.953
IGD	M	0.939	0.939	0.939	0.939
	B	0.964	0.964	0.964	0.964
	Avg.	0.954	0.954	0.954	0.954



Dataset 2:

Table 6: Class-wise Results for LR

Logistic Regression					
	Class	Accuracy	Precision	Recall	F-Measure
RAW Data	Alive	0.964	0.907	0.964	0.934
	Dead	0.436	0.681	0.436	0.532
	Avg.	0.885	0.873	0.885	0.874
Class Balancer	Alive	0.804	0.946	0.804	0.869
	Dead	0.740	0.400	0.740	0.519
	Avg.	0.795	0.864	0.795	0.817
Under-sample	Alive	0.803	0.947	0.803	0.869
	Dead	0.746	0.401	0.746	0.521
	Avg.	0.795	0.865	0.795	0.817
400% SMOTE	Alive	0.817	0.927	0.817	0.869
	Dead	0.635	0.380	0.635	0.475
	Avg.	0.790	0.845	0.790	0.810

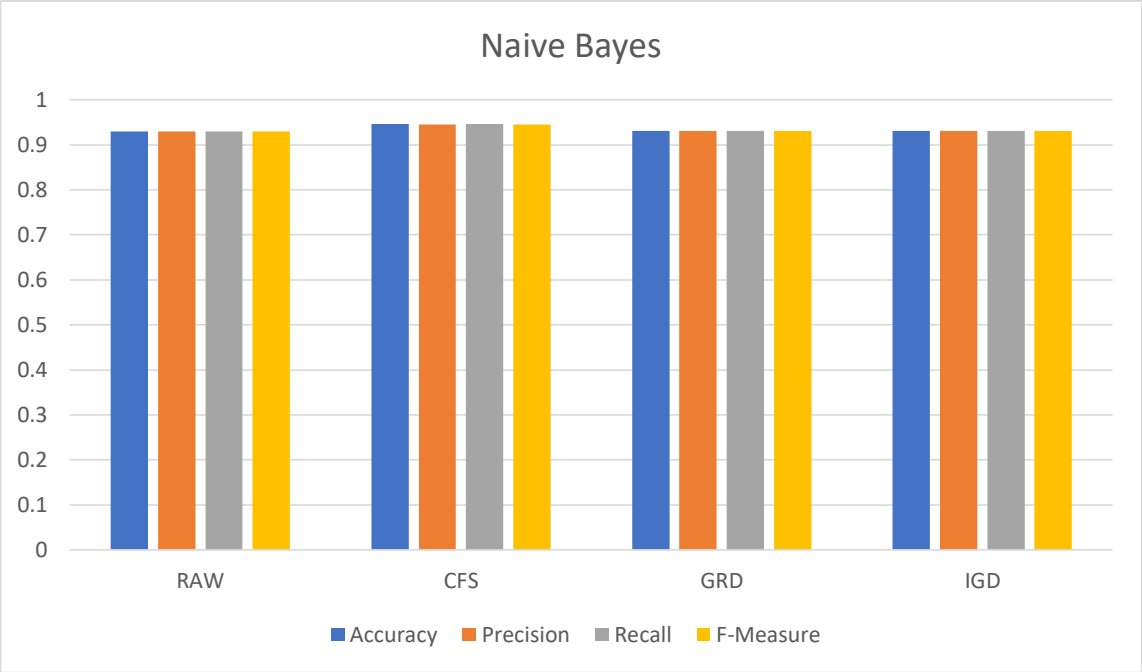


Naive Bayes Model

Dataset 1:

Table 7: Class-wise Results for NB

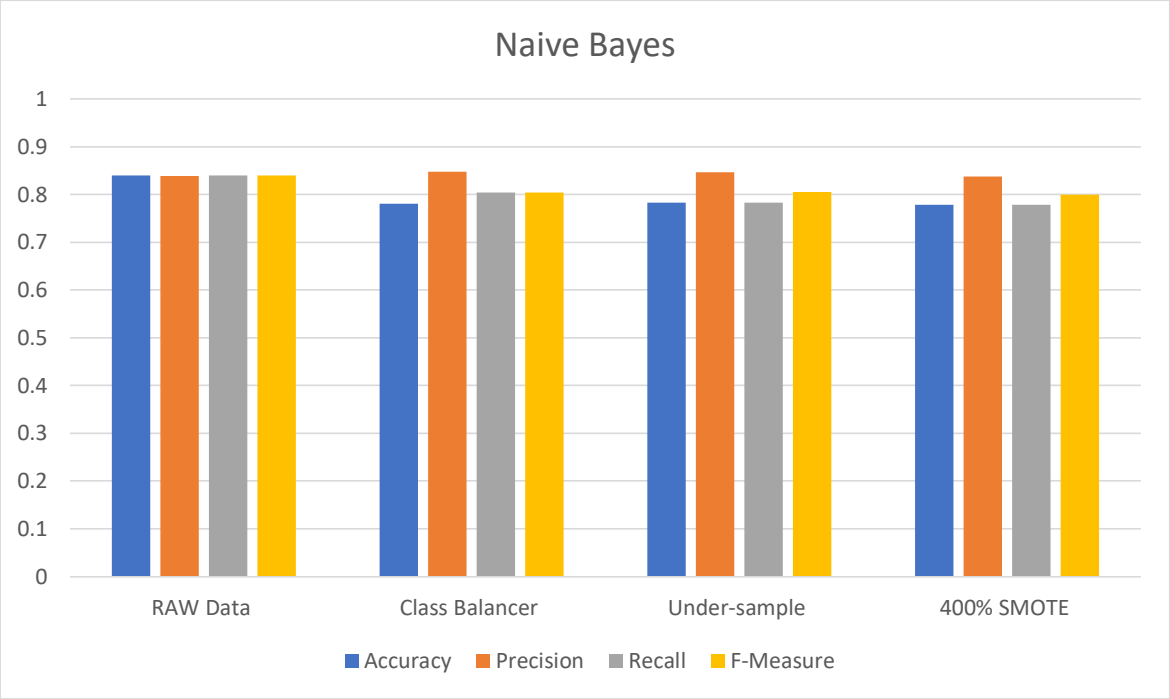
Naive Bayes					
	Class	Accuracy	Precision	Recall	F-Measure
RAW	M	0.896	0.913	0.896	0.905
	B	0.95	0.939	0.95	0.944
	Avg.	0.93	0.93	0.93	0.93
CFS	M	0.91	0.941	0.91	0.926
	B	0.966	0.948	0.966	0.957
	Avg.	0.946	0.945	0.946	0.945
GRD	M	0.896	0.918	0.869	0.907
	B	0.952	0.939	0.952	0.946
	Avg.	0.931	0.931	0.931	0.931
IGD	M	0.896	0.918	0.896	0.907
	B	0.952	0.939	0.952	0.946
	Avg.	0.931	0.931	0.931	0.931



Dataset 2:

Table 8: Class-wise Results for NB

Naive Bayes					
	Class	Accuracy	Precision	Recall	F-Measure
RAW Data	Alive	0.907	0.905	0.907	0.906
	Dead	0.459	0.466	0.459	0.462
	Avg.	0.840	0.839	0.840	0.840
Class Balancer	Alive	0.800	0.932	0.800	0.861
	Dead	0.669	0.371	0.669	0.477
	Avg.	0.781	0.848	0.804	0.804
Under-sample	Alive	0.806	0.929	0.806	0.863
	Dead	0.652	0.372	0.652	0.474
	Avg.	0.783	0.846	0.783	0.805
400% SMOTE	Alive	0.807	0.922	0.807	0.861
	Dead	0.613	0.359	0.613	0.453
	Avg.	0.778	0.838	0.778	0.800

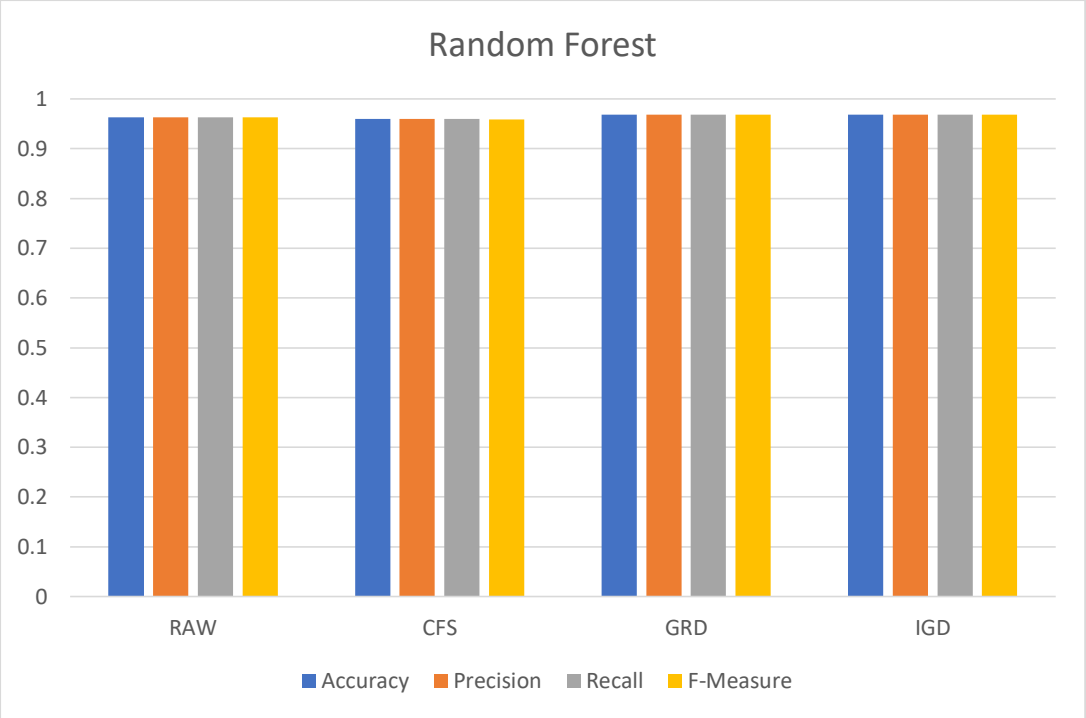


Random Forest Model

Dataset 1:

Table 9: Class-wise Results for RF

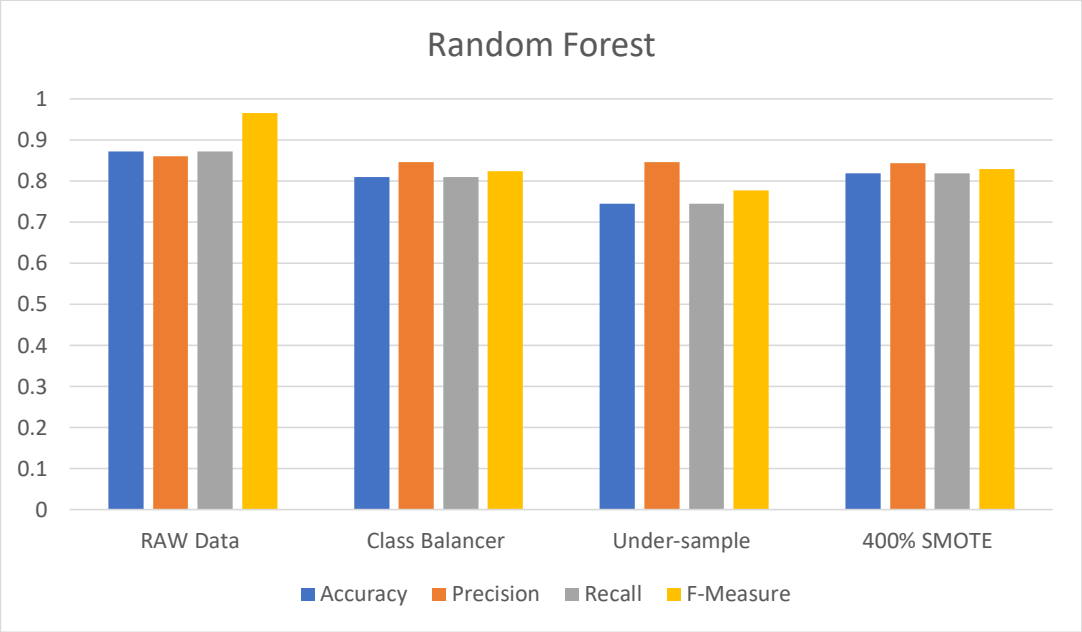
Random Forest					
	Class	Accuracy	Precision	Recall	F-Measure
RAW	M	0.934	0.966	0.934	0.95
	B	0.98	0.962	0.98	0.971
	Avg.	0.963	0.963	0.963	0.963
CFS	M	0.929	0.961	0.929	0.945
	B	0.978	0.959	0.978	0.968
	Avg.	0.96	0.96	0.96	0.959
GRD	M	0.943	0.971	0.943	0.957
	B	0.983	0.967	0.983	0.975
	Avg.	0.968	0.968	0.968	0.968
IGD	M	0.943	0.971	0.943	0.957
	B	0.983	0.967	0.983	0.975
	Avg.	0.968	0.968	0.968	0.968



Dataset 2:

Table 10: Class-wise Results for RF

Random Forest					
	Class	Accuracy	Precision	Recall	F-Measure
RAW Data	Alive	0.944	0.908	0.944	0.926
	Dead	0.459	0.593	0.459	0.517
	Avg.	0.872	0.861	0.872	0.965
Class Balancer	Alive	0.846	0.923	0.846	0.883
	Dead	0.602	0.408	0.602	0.487
	Avg.	0.810	0.846	0.810	0.824
Under-sample	Alive	0.752	0.936	0.752	0.834
	Dead	0.707	0.334	0.707	0.454
	Avg.	0.745	0.846	0.745	0.777
400% SMOTE	Alive	0.865	0.917	0.865	0.890
	Dead	0.558	0.421	0.558	0.450
	Avg.	0.819	0.843	0.819	0.829

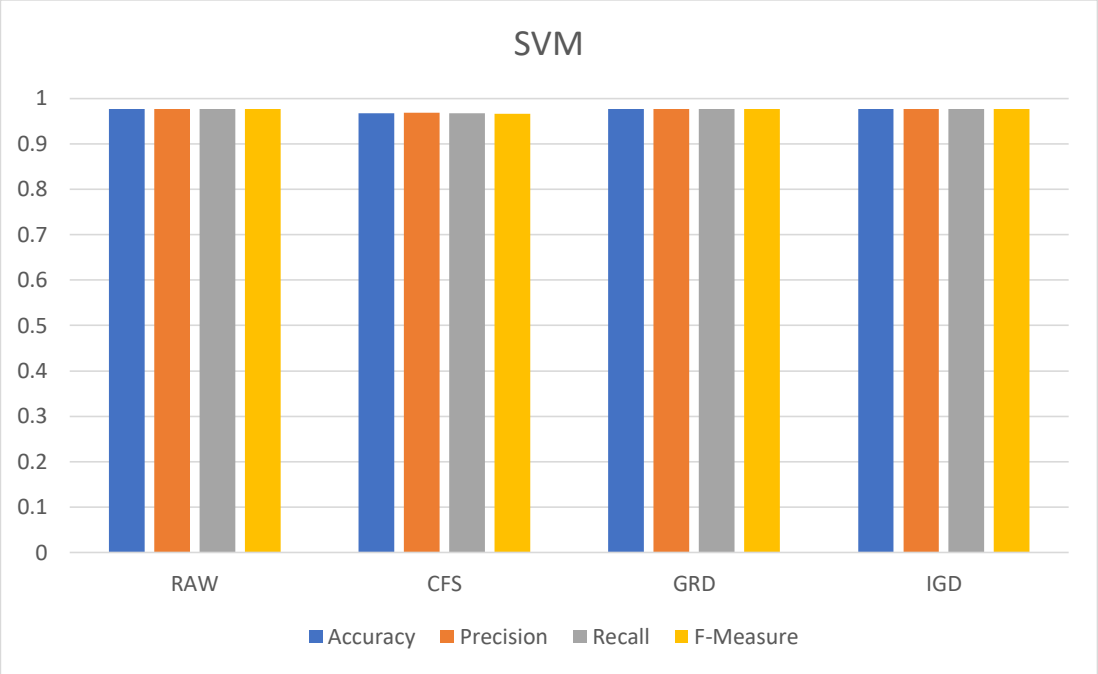


Support Vector Machine (SVM) Model:

Dataset 1:

Table 11: Class-wise Results for SVM

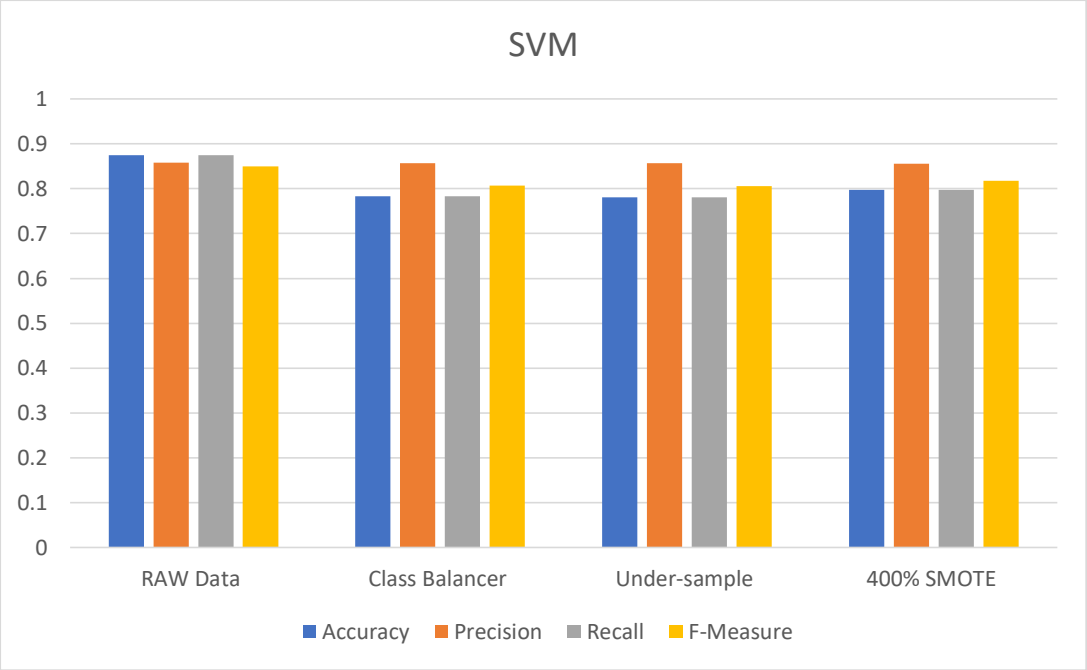
Support Vector Machine					
	Class	Accuracy	Precision	Recall	F-Measure
RAW	M	0.948	0.99	0.948	0.969
	B	0.994	0.97	0.994	0.982
	Avg.	0.977	0.977	0.977	0.977
CFS	M	0.915	0.995	0.915	0.953
	B	0.997	0.952	0.997	0.974
	Avg.	0.967	0.968	0.967	0.966
GRD	M	0.948	0.99	0.948	0.969
	B	0.994	0.97	0.994	0.982
	Avg.	0.977	0.977	0.977	0.977
IGD	M	0.948	0.99	0.948	0.969
	B	0.994	0.97	0.994	0.982
	Avg.	0.977	0.977	0.977	0.977



Dataset 2:

Table 12: Class-wise Results for SVM

SVM					
	Class	Accuracy	Precision	Recall	F-Measure
RAW Data	Alive	0.980	0.885	0.980	0.930
	Dead	0.276	0.704	0.276	0.397
	Avg.	0.874	0.858	0.874	0.850
Class Balancer	Alive	0.795	0.941	0.795	0.862
	Dead	0.718	0.381	0.718	0.498
	Avg.	0.783	0.857	0.783	0.807
Under-sample	Alive	0.793	0.941	0.793	0.860
	Dead	0.718	0.379	0.718	0.496
	Avg.	0.781	0.857	0.781	0.806
400% SMOTE	Alive	0.817	0.936	0.817	0.873
	Dead	0.685	0.397	0.685	0.503
	Avg.	0.797	0.856	0.797	0.817



Chapter 5 Result Analysis

In this section, the results of each algorithm will be discussed and compared to find the best algorithm for this specific dataset.

Dataset 1:

Table 13: Results for the algorithms in different feature selection methods

Algorithm	Accuracy	Precision	Recall	F-Measure
RAW DATA				
J48	0.933	0.934	0.933	0.933
KNN	0.960	0.960	0.960	0.960
Logistic Regression	0.942	0.943	0.942	0.942
Naive Bayes	0.930	0.930	0.930	0.930
Random Forest	0.963	0.963	0.963	0.963
SVM	0.977	0.977	0.977	0.977
Info Gain Data (IGD)				
J48	0.935	0.936	0.935	0.935
KNN	0.953	0.952	0.953	0.952

Logistic Regression	0.954	0.954	0.954	0.954
Naive Bayes	0.931	0.931	0.931	0.931
Random Forest	0.968	0.968	0.968	0.968
SVM	0.977	0.977	0.977	0.977
CFS Selection Data (CFS)				
J48	0.940	0.941	0.940	0.940
KNN	0.958	0.958	0.958	0.958
Logistic Regression	0.968	0.968	0.968	0.968
Naive Bayes	0.946	0.945	0.946	0.945
Random Forest	0.960	0.960	0.960	0.959
SVM	0.967	0.968	0.967	0.966
Gain Ratio Data (GRD)				
J48	0.931	0.933	0.931	0.932
KNN	0.953	0.952	0.953	0.952
Logistic Regression	0.953	0.953	0.953	0.953
Naive Bayes	0.931	0.931	0.931	0.931
Random Forest	0.968	0.968	0.968	0.968
SVM	0.977	0.977	0.977	0.977

We implemented four different classification algorithm models in the different feature selection methods.

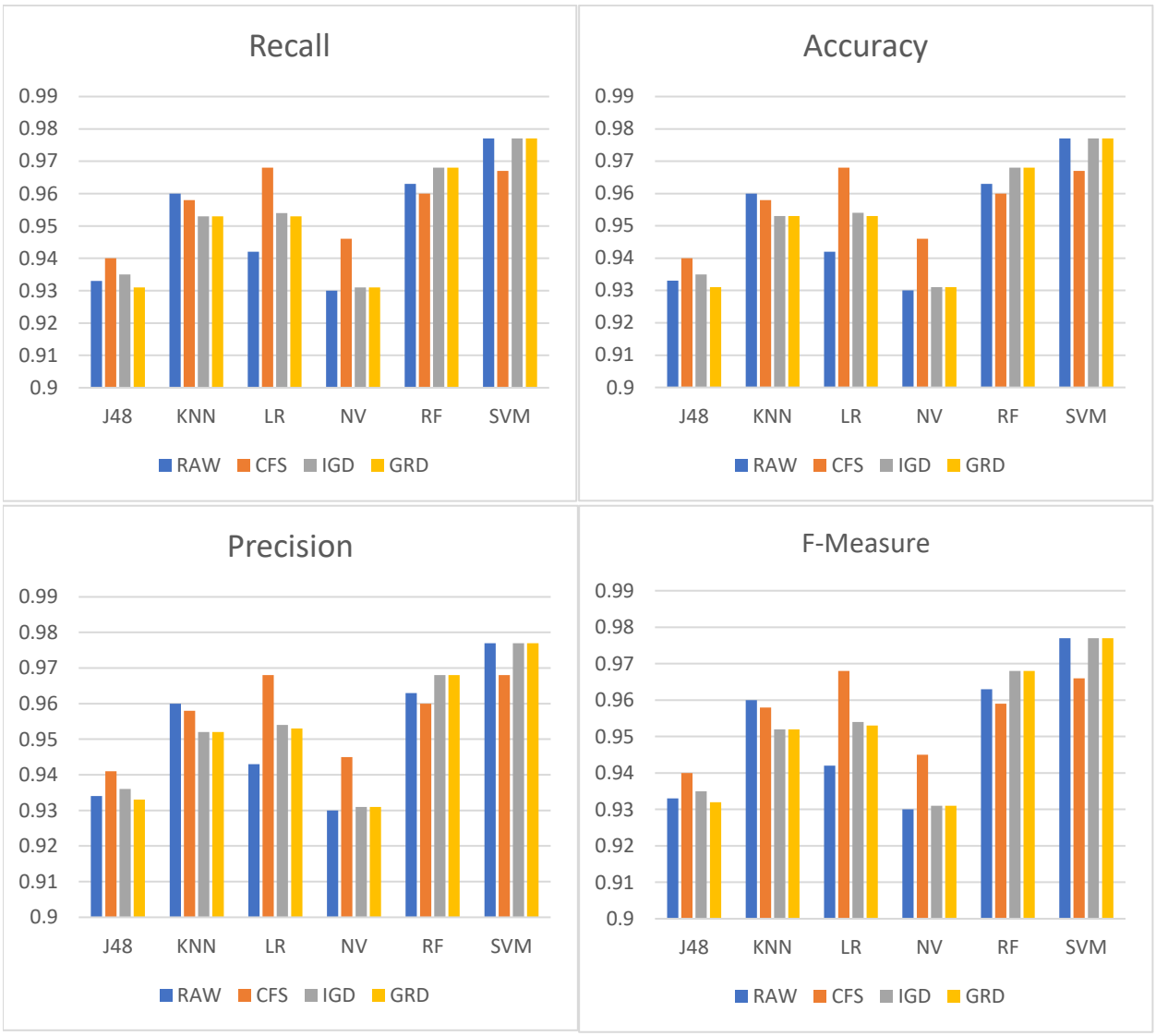


Figure 7: Accuracy, Precision, Recall & f-Measure for all the algorithms For dataset-1

Here we can see the classification results of a different model. We found that different models give different results in the feature selection method.

But for every class Support Vector Machine gives a consistent result in every model. So it is clear that SVM is the best classification method here.

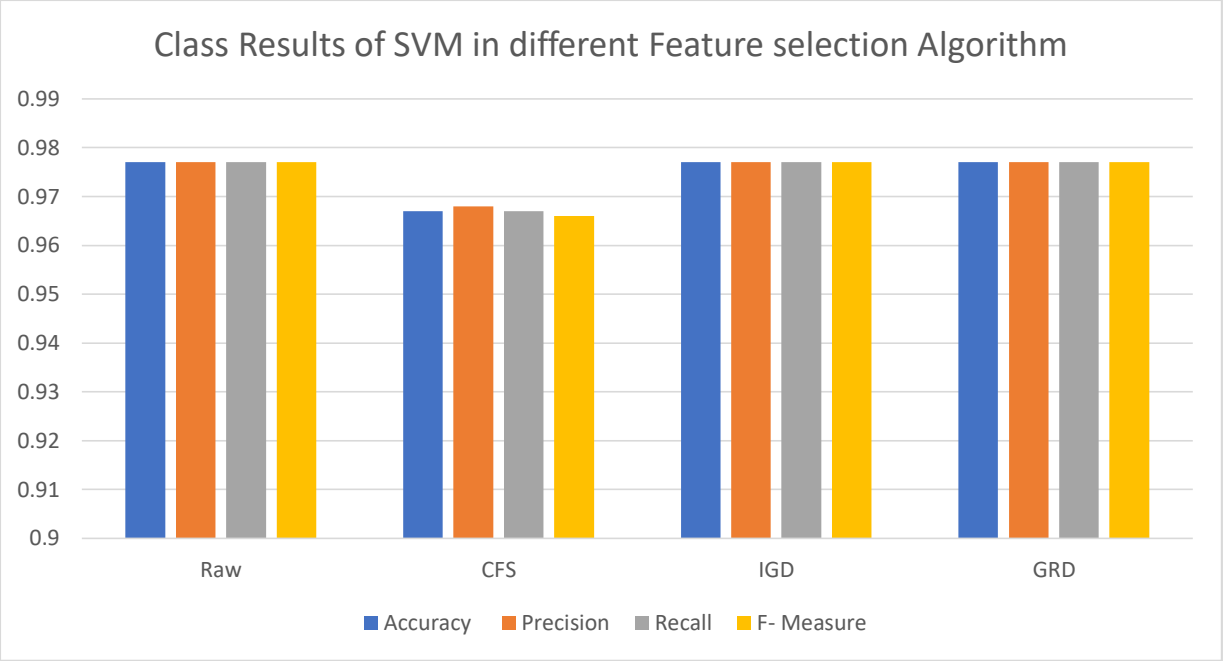


Figure 8: Class Results of SVM in different Feature Selection Algorithms

This chart shows the result of SVM for every class for different feature selection methods. But the best result is given when implemented on Info Gain Data (IGD) and Gain Ratio Data (GRD). Here both of the methods give a similar prediction result.

The Accuracy, Precision, Recall, and F- Measure value is 97.7% for both IGD and GRD.

Dataset 2

Table 14: Results for the algorithms for different class balance handling methods

Algorithm	Accuracy	Precision	Recall	F-Measure
RAW DATA				
J48	0.879	0.865	0.879	0.867
KNN	0.851	0.827	0.851	0.834
Logistic Regression	0.885	0.873	0.885	0.874
Naive Bayes	0.840	0.839	0.840	0.840
Random Forest	0.872	0.861	0.872	0.965
SVM	0.874	0.858	0.874	0.850
Class Balancer				
J48	0.778	0.846	0.778	0.802
KNN	0.776	0.828	0.776	0.796
Logistic Regression	0.795	0.864	0.795	0.817
Naive Bayes	0.781	0.848	0.804	0.804
Random Forest	0.810	0.846	0.810	0.824
SVM	0.783	0.857	0.783	0.807
Under-sample				
J48	0.806	0.863	0.806	0.825
KNN	0.759	0.832	0.759	0.785
Logistic Regression	0.795	0.865	0.795	0.817
Naive Bayes	0.783	0.846	0.783	0.805
Random Forest	0.745	0.846	0.745	0.777
SVM	0.781	0.857	0.781	0.806
400% SMOTE				
J48	0.825	0.854	0.825	0.836
KNN	0.781	0.825	0.781	0.798
Logistic Regression	0.790	0.845	0.790	0.810
Naive Bayes	0.778	0.838	0.778	0.800
Random Forest	0.819	0.843	0.819	0.829
SVM	0.797	0.856	0.797	0.817

In the case of dataset 2, we did not implement the different classification algorithm models in the different feature selection methods/models as we did not reduce any of the attributes by feature selection.

Here we implemented some imbalance handling to generalize the classes for every model.

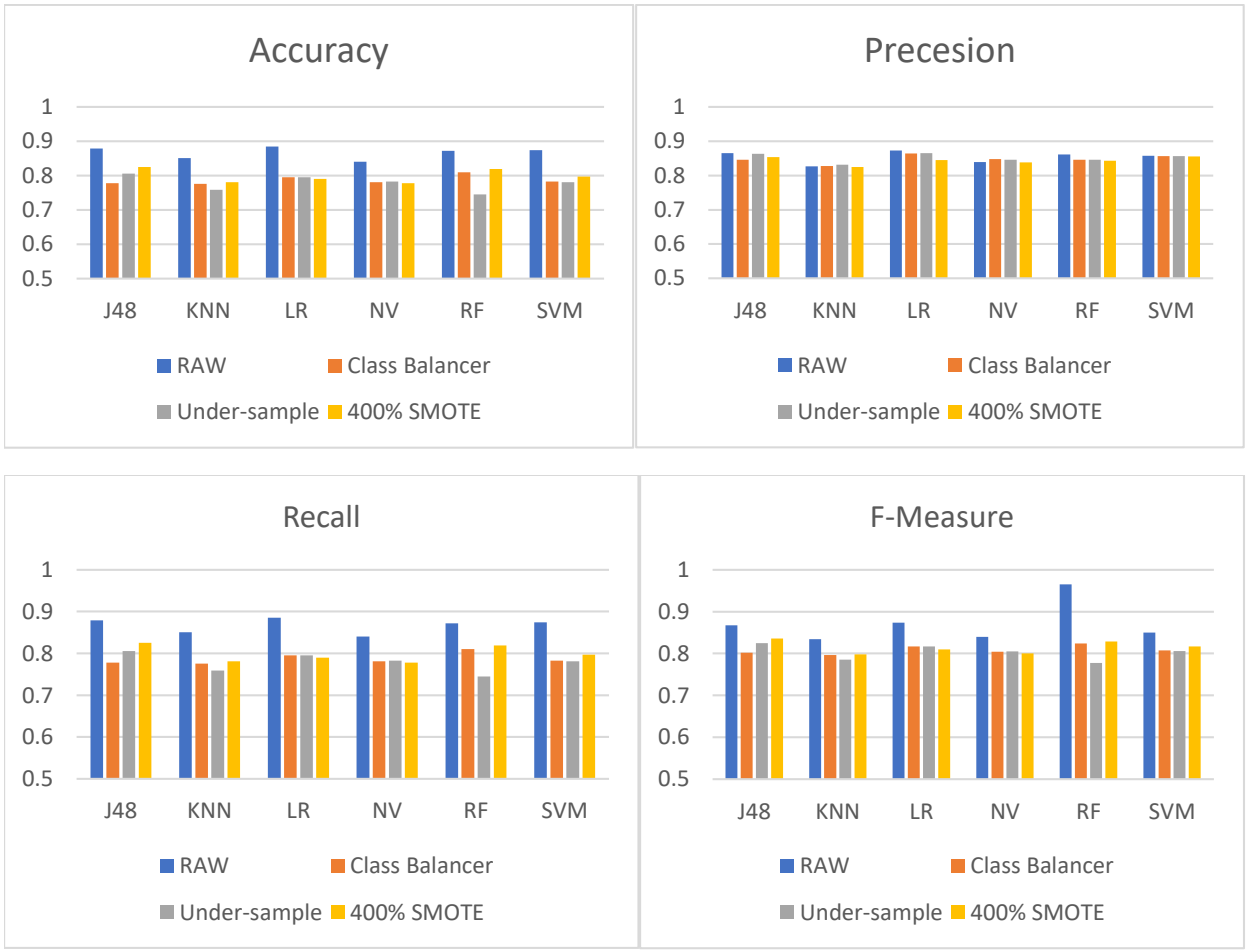


Figure 9: Accuracy, Precision, Recall & f-Measure for all the algorithms for dataset-2

In this dataset, we got different classification algorithm that works best with different imbalance handling method.

From the chart we can see here LR (Logistic Regression) gives a better result on average. But the result was not very balanced in all the classes. While we got a constant result from J48 (Decision Tree) in every case. And the prediction results of every class were quite generalized and balanced than the others.

The best result given by j48 is when it is implemented on under-sampling and 400% SMOTE.

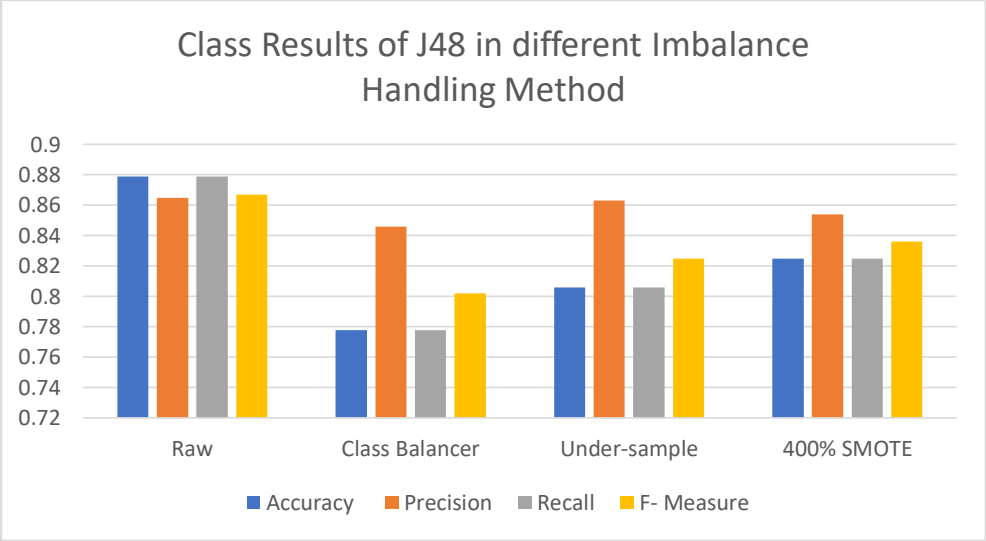


Figure 10: Class Results of J48 in different Feature Selection Algorithms

Here Under sampling in J48 gives a value of 80.6% (Accuracy), 86.3% (Precision), 80.6% (Recall), and 82.5% (F- Measure).

And 400% SMOTE gives a value of 82.5% (Accuracy), 85.4%(Precision), 82.5%(Recall), and 83.6% (F- Measure).

Figure 11: Class Balancing of J48

Though it looks like 400% SMOTE is the winner here as the average is quite higher. But the under-sampling has a generalized prediction value for both classes.

Chapter- 6 Conclusion and Future Works

After analyzing the results and evaluating the models we got the best algorithm on which we can rely in case of cancer prediction. The algorithm is SVM or Support Vector Machine algorithm which predicts 97.7% accurately. We got better results than previous works related to this issue. So, we will like to propose an SVM machine learning algorithm in case of cancer prediction.

And in the area of prediction of survival rate, we would like to propose the J48 algorithm which is a Decision tree algorithm. This model predicts almost 83% accurately.

Our work features better accuracy in the case of both prognosis and survival rate prediction. If we can predict the survival rate just after the prognosis prediction the treatment procedure will be very specific and well-decorated. In many cases starting the proper treatment for cancer patients takes a long way.

Despite having a small dataset the proposed model has scored surprisingly well without any overfitting issues.
In the future, we want to work on a large amount of data to get more feasible outcomes. There is a scope to use more advanced technology like Deep Learning to enhance the performance.

References

[1] WHO, "Cancer," WHO, 3 2 2022. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cancer#:~:text=Key%20facts,and%20rectum%20and%20prostate%20cancers. [Accessed 08 11 2022].

[2] UCI, "Breast Cancer Wisconsin (Diagnostic) Data Set," UCI, 1 11 1995. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic). [Accessed 8 11 2022].

[3] IEEE, "IEEEdataport," IEEE, 18 1 2019. [Online]. Available: https://ieee-dataport.org/open-access/seer-breast-cancer-data. [Accessed 8 11 2022].

[4] J. F. M. M. L. S. M. L. M. S. M. M. P. J. D. P. B. B. M. P. Hyuna Sung Ph.D., "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," *ACS Journals,* vol. 71, no. 3, pp. 209-249, 2021.

[5] T. Saba, "Recent advancement in cancer detection using machine learning: Systematic survey of decades, comparisons and challenges," *Journal of Infection and Public Health,* vol. 13, no. 9, pp. 1274-1289, 2020.

[6] M. A. T. Q. N. Y. A. O. E. Ali Bou Nassif*, "Breast cancer detection using artificial intelligence techniques: A systematic literature review," in *Artificial Intelligence in Medicine*, UAE, 2022.

[7] D. S. W. Joseph A. Cruz, "Applications of Machine Learning in Cancer Prediction and Prognosis," *SAGE,* no. January - December 2006, 2017.

[8] P. B. A. K. A. T. Parul Tiwari, "Breast Cancer Survival Prediction Using Machine Learning," in *Computational Intelligence in Oncology*, Singapore, Springer Nature Singapore Pte Ltd., 2022, pp. 143-158.

[9] M. N. I. a. A. K. Md. Toukir Ahmed, "Analysis of Wisconsin Breast Cancer original dataset using data," *Journal Binet,* vol. 9, no. 2, pp. 665-672., 2020.

[10] S. Sharma, A. Aggarwal, and T. Choudhury, "Breast Cancer Detection Using Machine Learning Algorithms," in *2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*, Belgaum, India., 2019.

[11] S. E. F. K. A. E. H. B. R. A. A. O. D. Mohammed Amine Naji, "Machine Learning Algorithms For Breast Cancer Prediction And Diagnosis," in *Procedia Computer Science*, Sciencedirect, 2021, pp. 487-492.

[12] Division of Cancer Prevention and Control, Centers for Disease Control and Prevention, "Breast Cancer," CDC, 26 9 2022. [Online]. Available: https://www.cdc.gov/cancer/breast/basic_info/what-is-breast-cancer.htm. [Accessed 8 11 2022].

[13] Cleveland Clinic, "Breast cancer," Cleveland Clinic, 21 1 2022. [Online]. Available: https://my.clevelandclinic.org/health/diseases/3986-breast-cancer. [Accessed 8 11 2022].

[14] S. H. Momenimovahed Z, "Epidemiological characteristics of and risk factors for breast cancer in the world," in *Breast Cancer*, Dove Med Press, 2019, pp. 151-164.

[15] M. F. S. Eva Singletary, "Rating the Risk Factors for Breast Cancer," in *ANNALS OF SURGERY*, Houston, Texas, Department of Surgical Oncology, The University of Texas, p. 474–482.

[16] N. S. F. D. C. W. Thomas G. Dietterich, "A Summary of Machine Learning Papers From IJCAI-85," in *Ninth International Joint Conference on Artificial Intelligence*, USA, 1985.

[17] V. C. B. MD, Thoracic Radiology: Noninvasive Diagnostic Imaging, Murray & Nadel's Textbook of Respiratory Medicine, 2022.

[18] M. Alloghani, "A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science," in *Supervised and Unsupervised Learning for Data Science*, 2020, pp. 3-21.

[19] G. S. D. R. B. M. M. Dr. Neeraj Bhargava, "International Journal of Advanced Research in," *International Journal of Advanced Research in Computer Science and Software Engineering,* vol. 3, no. 6, pp. 1114-1119, 2013.

[20] P.-N. T. Michael Steinbach, "kNN: k-Nearest Neighbors," in *The Top Ten Algorithms in Data Mining*, Chapman and Hall/CRC, 2009, p. 12.

[21] M. P. LaValley, "Logistic Regression," in *Circulation*, Lippincott Williams & Wilkins, 2008, pp. 2395 - 2399.

[22] D. A. M.Schnyer, "Chapter 6 - Support vector machine," in *Machine Learning Methods and Applications to Brain Disorders*, Department of Psychology, the University of Texas at Austin, Austin, TX, United States, 2019, pp. 101-121.

[23] G. I. Webb, "Naïve Bayes," in *Encyclopedia of Machine Learning and Data Mining*, 2017.

[24] S. J. Rigatti, "Random Forest," *Journal of Insurance Medicine,* vol. 47, no. 1, pp. 31-39, 2017.

[25] J. Heaton, "An empirical analysis of feature engineering for predictive modeling," in *IEEE*, Norfolk, VA, USA, 2016.

[26] H. Patel, "Towards Data Science," 30 08 2021. [Online]. Available: https://towardsdatascience.com/what-is-feature-engineering-importance-tools-and-techniques-for-machine-learning-2080b0269f10.

[27] J. Y. Zhou ZH, "Medical diagnosis with C4.5 Rule preceded by artificial neural network ensemble," *IEEE Trans Inf Technol Biomed,* vol. 7, no. 1, pp. 37-42, 2003.

[28] M. L. J. B. H. T. S. P. L. a. J. H. Lundin, "Artificial neural networks applied to survival prediction in breast cancer," in *Oncology*, 1999.

[29] D. W. G. a. K. A. Delen, "Predicting breast cancer survivability: a comparison of three data mining methods," in*. Artificial intelligence in medicine*, 2005, p. 113–127.

[30] J. M. a. D. W. Jiji, "An Efficient CBIR Approach for Diagnosing the Stages of Breast Cancer Using KNN Classifier.," *Bonfring International Journal of Advances in Image Processing,* vol. 2, no. 1, pp. 01-05, 2012.

[31] L. Liu, "Research on logistic regression algorithm of breast cancer diagnose data by machine learning," *2018 International Conference on Robots & Intelligent Systems (ICRIS),* p. 157–160, 2018.

[32] N. G. S. M. H. G. F. Q. M. B. Hakim El Massari1*, "An Ontological Model based on Machine Learning for Predicting Breast Cancer," *International Journal of Advanced Computer Science and Applications,* vol. 13, no. 7, pp. 108-115, 2022.

[33] T. Y. K. H. H. K. Zhao M, "Machine Learning With K-Means Dimensional Reduction for Predicting Survival Outcomes in Patients With Breast Cancer," *Cancer Inform,* 2018.

[34] T. D. J. Dr.B.Santhosh Kumar, "Breast Cancer Prediction Using Machine Learning Algorithms," *International Journal of Advanced Science and Technology,* vol. 29, no. 3, pp. 7819 - 7828, 2020.

[35] P. J. R. D. T. A. H. S. M. D. Rezaianzadeh A, "Survival analysis of 1148 women diagnosed with breast cancer in Southern Iran," *BMC Cancer,* vol. 9, p. 168, 2009.

[36] A. Y. Y. U. S. N. S. K. Al Zahrani AM, "Quality of life of women with breast cancer undergoing treatment and follow-up at King Salman Armed Forces Hospital in Tabuk, Saudi Arabia," in *Breast Cancer*, Saudi Arabia, Dove Med Press), 2019, pp. 199-208.

[37] K. &. R. C. Juneja, "An improved weighted decision tree approach for breast cancer prediction.," *International Journal of Information Technology.,* 2018.

[38] A. &. E.-M. S. Azar, "Decision tree classifiers for automated medical diagnosis.," in *Neural Computing and Applications*, 2013, pp. 2387-2403.

[39] M. &. H. M. &. I. H. &. H. M. M. &. H. M. &. K. M. N. slam, "Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques.," in *SN Computer Science*, 2020.

[40] A. v. M. M. H. M. Moncada-Torres, "Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival," *Sci Rep,* 2021.

[41] [Online]. Available: https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic).

[42] [Online]. Available: https://ieee-dataport.org/open-access/seer-breast-cancer-data.

[43] NCI, "Stages of cancer," NCI, 14 10 2022. [Online]. Available: https://www.cancer.gov/about-cancer/diagnosis-staging/staging.

[44] Cancer Research UK, "cANCER sTAGING," Stages of cancer, 7 7 2020. [Online]. Available: https://www.cancerresearchuk.org/about-cancer/what-is-cancer/stages-of-cancer.

[45] E. R.-F. J. B. F. e. a. Rakha, "Breast cancer prognostic classification in the molecular era: the role of histological grade.," *BMC,* 2010.

[46] NHS, "What do cancer stages and grades mean? " NHS, 16 12 2021. [Online]. Available: https://www.nhs.uk/common-health-questions/operations-tests-and-procedures/what-do-cancer-stages-and-grades-mean/#:~:text=The%20grade%20describes%20the%20appearance,and%20the%20best%20treatment%20options..

[47] A. C. Society, Breast Cancer Facts & figures 2019-20, Atalanta: American Cancer Society, 2020.

[48] Webmd, "Types of Breast Cancer," [Online]. Available: https://www.webmd.com/breast-cancer/breast-cancer-types-er-positive-her2-positive#.

[49] S. M. K. N. S. F. Trabert B, "Progesterone and Breast Cancer," *Endocr Rev. 2020,* p. 320–344., 2020.

[50] S. A. Frank, Dynamics of Cancer, Princeton University Press, 2007.

[51] P. C. L. C. e. a. Carey LA, "Race, Breast Cancer Subtypes, and Survival in the Carolina Breast Cancer Study. JAMA.," *JAMA,* 2006.

[52] S. A. Frank, Dynamics of Cancer: Incidence, Inheritance, and Evolution, Princeton, Princeton University Press, 2018.

[53] D. Berrar, "Cross-validation," *Encyclopedia of Bioinformatics and Computational Biology,* vol. 1, p. 542–545, 2018.

[54] V. P. M. E. A. e. a. Singh, "Impact of train/test sample regimen on performance estimate stability of machine learning in cardiovascular imaging," *Sci Rep 11,* vol. 11, 2021.

[55] K. F. Erickson BJ, "Magician's Corner: 9. Performance Metrics for Machine Learning Models," in *Radiol Artif Intell*, 2021.

[56] Z. C. Yixuan Li, "Performance Evaluation of Machine Learning Methods for Breast Cancer Prediction," *Applied and Computational Mathematics,* vol. 7, no. 4, pp. . 212-216, 2018.