Meta – Analysis of Genomic Data in Endometrial Cancer

A dissertation submitted for the Degree of Master of Computer Science

T A D S U Siriwardana

University of Colombo School of Computing

2018

Declaration

The thesis is my original work and has not been submitted previously for a degree at this or any other university/institute.

To the best of my knowledge it does not contain any material published or written by another person, except as acknowledged in the text.

Student Name: T A D S U Siriwardana

Registration Number: 2015/MCS/072

Index Number: 15440721

---------- ----------

Signature: Date:

This is to certify that this thesis is based on the work of

Ms. T A D S U Siriwardana

under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Certified by:

Supervisor Name: Mrs. M W A C R Wijesinghe

---------- ----------

Signature: Date:

Abstract

Endometrial Carcinoma (EC) is the most common gynecologic malignancy in the Asian Countries as well as other developing countries. The majority of cases are diagnosed when the carcinoma is confined to the uterus, leading to less than 1.5% of cancer deaths. Endometrial Cancers largely caused by hormonal imbalances. But the research done to identify the genomic level impact is limited. Among different types of mutations Copy Number Variation(CNV) has a significant effect in causing tumors. Studying the Copy Number Variation (CNV) in hormone responsive genes of endometrial cancer would help to understand the biological underpinning of endometrial cancer progression and to determine the treatment strategies.

In this study, we have proposed an approach to identify the genomic therapies for endometrial cancer by predicting the gene alteration associated with hormone responsive genes. Patient's clinical data are analyzed and identified how the alterations of hormone responsive genes. In this study Classification is used to predict the correlation between patient data samples and classes by analyzing Patients data. Prediction model is built using Support Vector Machine (SVM) to discover correlation between patient clinical data and gene alterations.

In this study we used two approaches to build the prediction model. Predicting the variation of single gene and predicting the variation of gene profiles are two methodologies. Output of the study shows that, the proposed SVM model gives a strong and an accurate prediction for possible alteration of hormone responsive gene of a patient.

Acknowledgement

I would like to express my deepest sincere gratitude towards my supervisor, Senior Lecture Mrs. Rupika Wijesinghe for her proper guidance, supervision, helpful advice and sparing his valuable time from the very early stage of this research as well as giving me extraordinary experiences throughout the work and she provided me unflinching encouragement and support in various ways.

I would also like to thank Dr. Ruvan Weerasinghe Senior Lecture for sparing his valuable time and guide me on this work.

I offer my obeisance to the entire senior and junior lecturer panel and staff of UCSC for the help and inspiration they extended advice, supervision, and crucial contribution on this research.

It is pressure to convey my special thanks for my family and friends supported me to make this study success.

Finally, I would like to offer my heartiest gratitude to all the people whose names are not appeared but their untiring effort was very much crucial to make this study success throughout the work.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

- EC - Endometrial Cancer
- CNV - Copy Number Variation
- SNPs - Single Nucleotide Polymorphisms
- PCA - Principal Component Analysis
- SVM - Support Vector Machine
- UEC - Uterine Corpus Endometrial Carcinoma
- UC - Uterine Carcinosarcoma
- ROC - Receiver Operating Characteristic

CHAPTER 1

1. Introduction

1.1 Introduction

This study focuses on the application of data analytics concepts to develop an efficient and effective approach to identify clinically actionable genetic variants of Endometrial cancer for diagnosis and individualized therapy.

Cancer is a result of the abnormal growth of cells that have the ability to invade or spread to other parts of the human body. For cancer to develop, genes regulating cell growth and differentiation must be altered. These mutations are then maintained through subsequent cell divisions and are thus present in all cancerous cells. Cancer is fatal due to its ability to spread all over the human body. Cancer cells can spread locally by moving into nearby normal tissues. Cancer can also spread regionally, to nearby lymph nodes, tissues, or organs and it can spread to distant parts of the human body as well. This is called metastatic cancer [1].

Endometrial cancer is the eighth most common type of cancer in women with approximately 320 000 cases recorded in 2012, accounting for about 4.8% of all cancers in women [2]. Endometrial cancer is a cancer that arises from the uterus. This affects mainly postmenopausal women. The average age of women diagnosed with endometrial cancer is 60. A woman's hormone balance affects a part in the growth of most endometrial cancers. Many of the risk factors for endometrial cancer affect estrogen levels. There are different types of endometrial cancer and present study mainly focuses only on 2 types. They are Uterine Corpus Endometrial Carcinoma(UEC) and Uterine Carcinosarcoma(UC).

Endometrial carcinoma starts in the endometrium, the inner layer of tissue lining the uterus, while sarcoma begins in the outer layer of muscle of the uterus. It is a rare cancer, Uterine Carcinosarcoma makes up less than 5% of all uterine cancers. Approximately 35% of patients survive five years after diagnosis [3].

The notion of altered genes is a powerful simplifying concept that enables users to analyze complex data sets and to develop biological hypotheses regarding recurrently altered gene sets and biologic pathways.

This study focuses on analyzing and determining whether there is a relationship between the clinical data and CNV to identify severity of hormonal responsive genes based on copy number variation in endometrial Carcinoma. Detecting CNV of genes is a complex and time consuming process. This study proposes a new model for identifying genomic variations based on CNV using patient clinical data, as a replacement for above mentioned complex and time consuming methodology used in Biology.

1.2 Motivation

In today's world the number of cancer patients are being increased at a significant rate. The main cause for this tragedy can be identified as the improper modern life style of human. As a result, nowadays cancer which is a life threatening deadly disease can be identified as a normal disease among people.

Endometrial carcinoma (EC) is the leading cancer of the female genital tract in the Asian Countries as well as other developing countries. This become the fourth most common cancer among women after breast, lung, and colorectal cancer. Changes in the balance of female hormones in the body can cause the endometrial cancer.

Certain conditions or diseases lead to changes that affect the balance between the estrogen and progesterone levels in the body. These changes can result in the thickening of the uterine lining and a subsequent increased risk for cell abnormality and cancer. Hormonal risk factors include:

- Polycystic ovarian syndrome
- Endometrial polyps or other benign growths in the endometrium
- Hormone therapy with tamoxifen for breast cancer
- Ovarian tumors that release estrogen

Hence identifying hormonal responsiveness which is lead to alteration of genomes, will be a major advantage to recognize causes for cancer and to determine best treatments.

1.3 Aims and Objectives

The main objective of this research is to identify the genomic therapies for endometrial cancer by predicting the gene alteration associated with hormone responsive genes. To achieve this aim following goals are identified.

1. Identification of hormone responsive genes associated with endometrial cancer based on previous literature.
2. Exploration of cancer type specific Copy Number Variations(CNV) associated with hormone responsive genes in Uterine Corpus Endometrial Carcinoma(UEC) and Uterine Carcinosarcoma (UC).
3. Predicting the Copy Number Variation (CNV) associated with hormone responsive genes based on patient's clinical data.

1.4 Problem Definition

Recently large number of sequencing studies of primary and metastasis endometrial cancers have indicated that the difference in phenotype is reflected in distinct molecular subgroups, with further molecular sub-clustering.

The hormone-related cancers shared a quite different mechanism of carcinogenesis. Hormones include both endogenous and exogenous. Hormones drive cell proliferation and cause increased number of cell divisions and results random genetic errors. The genetic basis of endogenous hormone levels as an important risk factor for hormone-dependent tumors.

Endometrial Cancers largely caused by hormonal imbalances. But the research done to identify the genomic level impact is minimum. Among difficult types of mutations CNV has a significant effect in causing tumors. Studying the CNV in hormone responsive genes of endometrial cancer would help to understand the biological underpinning of endometrial cancer progression and to determine the treatment strategies.

1.5 Scope of the Project

The study focuses on analyzing the gene expression data of two different endometrial cancer types, Uterine Corpus Endometrial Carcinoma(UEC) and Uterine Carcinomasarcoma(UC). Patient data will be obtained from cBioportal. The cBioPortal for Cancer Genomics ([http://cbioportal.org||cBio

Portal]) is specifically designed to lower the barriers of access to the complex data sets and thereby accelerate the translation of genomic data into new biological insights, therapies, and clinical trials.

The Cancer Genome Atlas(TCGA) provisional has provided 548 patients records for Uterine Corpus Endometrial Carcinoma and 57 records for Uterine Carcinomasarcoma. For each patient's data about approximately 100 clinical attributes and mutations more than 5000 genes are recorded. Mutations include Copy Number Variations (CNV) and Single Nucleotide Polymorphism (SNPs). CNV data of important hormone responsive genes will be studied in this work.

CHAPTER 2

2. Literature Review

2.1 Introduction

This chapter details the literature review on methodologies of analyzing genomic data in different diseases. Here the study reviews the literature on Exploration of Cancer type specific mutations, Identification of hormones on specific genes, analyzing clinical data sets on genetic variations based on copy number variation (CNV) associated with Hormonal responsiveness of human endometrial carcinoma. Further the study reviews different types of methodologies to manipulate and analyze clinical data sets and various methods to gain knowledge from those manipulated data sets.

2.2 Endometrial Carcinoma (EC)

Endometrial carcinoma (EC) is the leading cancer of the female genital tract in most of the countries. It is the sixth most common cancer type in women worldwide. Out of all the cases, about 53% of endometrial cancer cases occurred in high developed countries [4]. The highest incidence of endometrial cancer was recorded in Northern America and Europe [5]; and the lowest was recorded in Africa and Asia.

Endometrial cancer begins in the layer of cells that form the lining (endometrium) of the uterus. It is sometimes called uterine cancer [6]. Endometrial cancer often produces symptoms at relatively early stages [5], so the disease is generally diagnosed early. As it frequently produces abnormal vaginal bleeding, which prompts women to see their doctors. If endometrial cancer is discovered early, removing the uterus surgically often cures endometrial cancer.

Genetic predictors of endometrial cancer risk that allow early detection of the disease are important for prevention and improved management strategies. Mutations in the mismatch repair (MMR) genes TP53, MLH1, MSH2, MSH6, PMS1 and PMS2 are known to confer increased risk in a proportion of endometrial cancer cases, and the mutation range includes copy number variants (CNVs) and Single nucleotide polymorphisms(SNPs).

2.3 Hormonal Responsiveness

The uterine endometrium is largely sensitive to hormones. The best studied and most relevant steroids that participate in endometrial cellular processes are the estrogens and progesterone [7]. These hormones tightly regulate the complex functioning of the female reproductive tract. Both hormones are intimately involved in controlling the growth, development, and remodeling of reproductive tissues as well as for the cyclic changes that occur during the menstrual cycle. These steroids function by binding to nuclear receptor proteins that act as transcription factors to modulate the expression of genes.

Estrogen or estrogen-related molecules bind specifically to the estrogen receptor (ER) whilst progesterone or progesterone-related molecules bind specifically to the progesterone receptor (PR). Both ER and PR are members of the nuclear receptor superfamily and therefore they share the structural homology and the key functional domains that are common to this group [8].

Normal endometrial function requires a balance of progesterone (PR) and estrogen (ER) effects. An imbalance caused by increased ER action or decreased PR action can cause in abnormal endometrial proliferation and finally, endometrial adenocarcinoma, the common cancer in women [9].

Previous studies performed some techniques to identify which genes impact to estrogen progesterone normal routine. A microarray technology is a very valuable method for the prediction of hormone-responsive activities in various gene expressions [10]. Gene expression microarray studies have been performed to identify the estrogen-responsive genes in breast cancer. However, the effectiveness of expression microarray analysis is decreased by technical issues such as relatively high levels of noise and low sensitivity.

In addition to the estrogen-responsive genes identified by many other genes have been found to be estrogen-regulated, using techniques such as RT-PCR and Northern and Western blot analyses. However, among them, only a small number have been shown to possess functional EREs within the transcription regulatory region.

Furthermore, transcriptome analyses such as RNA-sequencing and CAGE sequencing56, mapping of ER binding sites (ERBSs) is being performed by ChIP-chip or ChIP sequencing (ChIP-seq) analysis, which is a deep-sequencing technique for ChIP-derived DNA fragments [11]. Integrated studies of transcriptomes and transcription factor binding sites can provide useful

information for ER-mediated gene regulation in a genome-wide manner. In particular, the development of next-generation sequencing has allowed us to obtain DNA sequence information much more rapidly, and from across the genome [11].

By performing different approaches previous studies found that estrogen and progesterone receptors such as ESR1 ESRRG PGR genes highly effect to the normal hormone routine.

The duplications, deletions or amplifications of genes on a chromosomal level, can be lead to many disorders. This is known as Copy Number Variations. The alterations of above identified hormone responsiveness gene can be a result of endometrial cancer.

2.4 Copy Number Variation(CNV)

Copy number variation (CNV) is a relatively new field in genomics and it is defined as a phenomenon in which part of the genome are repeated and the number of repeats in the genome diverse between individuals in the human population. Copy number variation is a type of structural variation: precisely, it is a type of duplication or deletion event that affects a significant number of base pairs.

CNVs can be generally categorized into two main groups: short repeats and long repeats. Short repeats include mainly bi-nucleotide repeats and tri-nucleotide repeats [12]. Long repeats include repeats of entire genes. This classification based on size of the repeat is the most explicit type of classification as size is an important factor in examining the types of mechanisms that most likely gave rise to the repeats, hence the likely effects of these repeats on phenotype.

It was normally thought that genes were always existing in two copies in a genome. However, recent findings have discovered that large segments of DNA, ranging in size from thousands to millions of DNA bases, can vary in copy-number. Such copy number variations (or CNVs) can surround genes leading to dosage imbalances. For example, genes that were thought to always occur in two copies per genome have now been found to sometimes be present in one, three, or more than three copies.

2.4.1 Copy Number Analysis(CNA)

Copy number analysis generally refers to the process of analyzing data formed by a test for DNA copy number variation in patient's sample. These type of analysis supports identify chromosomal copy number variation which may cause or may increase risks of various serious disorders. Copy

number variation can be detected with various types of tests such as high-resolution array-based tests, SNP array technologies and high resolution microarrays that include copy number probes as well an SNPs [13]. Array-based methodologies for copy number analysis offer reliable, efficient techniques for large-scale analysis [14]. Data analysis for an array-based DNA copy number test can be very challenging though due to very high volume of data that come out of an array platform.

BAC (Bacterial Artificial Chromosome) arrays were historically the first microarray platform to be used for DNA copy number analysis. This platform is used to identify gross deletions or amplifications in DNA. Such irregularities for example are common in cancer and can be used for diagnosis of many progressive disorders. Data produced by such platforms are usually low to medium resolution in terms of genome coverage. Usually, log-ratio measurements are produced by this technology to represent deviation of patient's copy number state from normal. Such measurements then are studied and those that considerably differ from zero value are announced to represent a part of a chromosome with an irregularity. Positive log-ratios indicate a region of DNA copy number gain and negative log-ratio values mark a region of DNA copy number loss. Even a single data point can be declared an indication of a copy number gain or a copy number loss in BAC arrays [13].

2.4.2 Single Nucleotide Polymorphism(SNP)

Single nucleotide polymorphisms, commonly called SNPs are the most common type of genetic variation among human genes. Each SNP denotes an alteration in a single DNA building block, called a nucleotide.

Most SNPs have no effect on health or development. Some of these genetic differences, however, have proven to be very important in the study of human health. Researchers have found SNPs that may help predict an individual’s response to certain drugs, susceptibility to environmental factors such as toxins, and risk of developing particular diseases [15].

Classifying tumors and identifying therapeutic targets requires a description of the genetic changes underlying cancer. Single nucleotide polymorphism (SNP) arrays provide a high-resolution platform for describing several types of genetic changes simultaneously. With the resolution of these arrays increasing exponentially, they are becoming increasingly powerful tools for describing the genetic events underlying cancer. SNP arrays are an ideal platform for identifying both somatic and germline genetic variants that lead to cancer. They provide a basis for

DNA-based cancer classification and help to define the genes being modulated, improving understanding of cancer genesis and potential therapeutic targets.

2.4.3 Why are CNVs important?

Differences in the DNA sequence of our genomes contribute to our uniqueness. These changes influence most traits including susceptibility to disease. It was thought that single nucleotide changes (called SNPs) in DNA were the most prevalent and important form of genetic variation. The current studies reveal that CNVs comprise at least three times the total nucleotide content of SNPs. Since CNVs often encompass genes, they may have important roles both in human disease and drug response. Understanding the mechanisms of CNV formation may also help us better understand human genome evolution.

2.5 Data sources

2.5.1 The cBio Cancer Genomics Portal

The cBio Cancer Genomics Portal ([http://cbioportal.org||cBio Portal]) is an open-access resource for interactive survey of multidimensional cancer genomics data sets, presently providing access to data from more than 5,000 tumor samples from 31 cancer studies [16]. The cBio Cancer Genomics Portal ([http://cbioportal.org||cBio Portal]), established at Memorial Sloan-Kettering Cancer Center (MSKCC) [16], was specifically developed to discourse the unique data integration issues posed by large-scale cancer genomics projects and to make the raw data generated by large-scale cancer genomic projects more easily and straightly available to the whole cancer research community. Statistical analysis, charts, graphs, tabular forms on different type of cancer data are available on cBio portal (Figure 2.1)

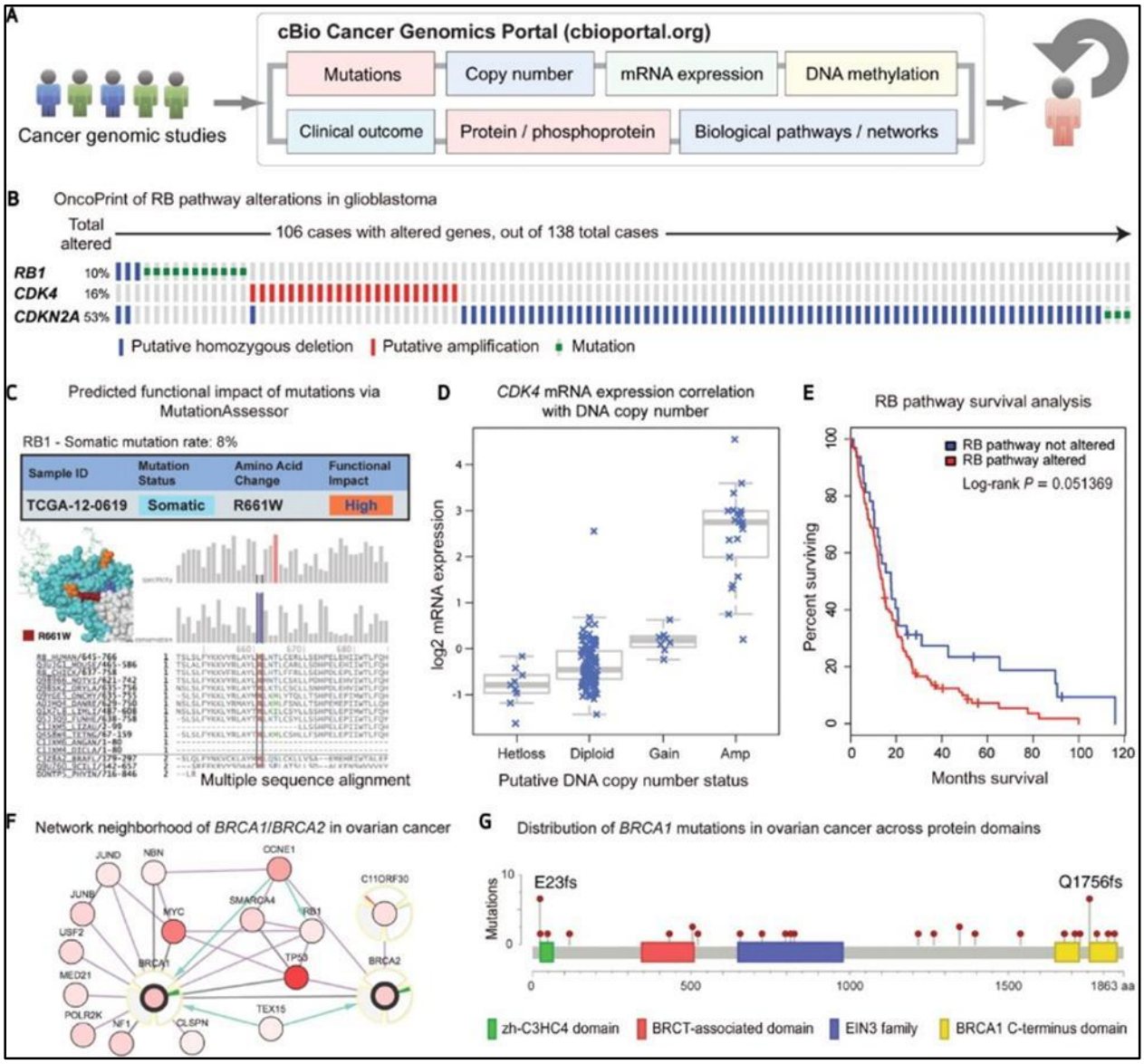

Figure 2.1:Interface of cBio portal

The cBio portal helps to access cancer genomic data sets for the entire biomedical community. It Provides a simple yet flexible interface to integrated data sets, intuitive visualization options, and a programmatic web interface, all of which can aid researchers in translating cancer genomic data into biologic insights and potential clinical applications. By incorporating multiple genomic data types and lowering the barrier to access, the portal enables researchers to more easily query genomic data, test hypotheses regarding genetic alterations in cancer, and place genomic data in the context of prior biologic knowledge. The cBio portal complements existing tools, including the TCGA and ICGC [17] data portals, the IGV [18], and IntOGen [4] by providing a

distinctive focus on analyzing discrete genomic events across integrated data types, ease of use, support for exploratory data analysis, and interactive network analysis.

2.6 Existing Approaches

The neural network methods can be easily used to accommodate nonlinear features of the gene expression data. It can produce continuous variables instead of discrete class labels. Neural network methods map input data into different classes. Which is quite useful in predicting the medical indicator level rather than classifying the samples into binary categories. The neural network method is a complex method and takes considerable time to master.

For many reasons such as inadequate resolution, image corruption, dirt or scratches on the slides or experimental error during the laboratory process, data can be lost in gene expression. So finding of missing value becomes an important preprocessing step. Since repeating the experiment is costly and time consuming, many algorithms have been introduced to recover loss data. Singular Value Decomposition based method (SVD impute) and weighted k-nearest neighbor's imputation (KNN impute) are used for estimating the missing value [19]. These two algorithms work well on deferent conditions. KNN impute shows better results on non-time series or noisy time series data, whereas, SVD impute works well on time series data with low noise levels. It has been shown that KNN impute offers a more solid methodology for missing value estimation than the SVD based method.

Furey et.al [20], introduced a gene classification method. Support Vector Machines have been used for gene data classification and produces good results when analyzing microarray expression data for genes from several tissue or cell types.

A method of functionally classifying genes introduced by Brown et al. [21] it is used to classify gene expression data from DNA microarray hybridization experiments using Support Vector Machine.

Support Vector Machine are widely used in gene expression analysis given to their many attractive features. One of the main reasons for adopting SVMs in gene expression analysis is their flexibility. They are capable of handling large feature spaces, and have the ability to identify outliers as well.

In most studies related to breast cancer, SVM classifier for breast cancer prediction are only based on the RBF kernel function and the Wisconsin dataset [22]. RBF is the most widely used kernel function in SVM but there are other the prediction performances obtained using other different popular kernel functions which have not yet been fully explored. The publicly available Wisconsin dataset is too small to effectively validate the performance of SVM for breast cancer prediction.

Furthermore, the prediction (or classification) accuracy is generally the primary evaluation metric used to measure the performance of prediction models. However, the collected dataset for specific cancer prediction is usually classified as a class imbalance problem. This causes a problem when examining the accuracy of the prediction model because the error of incorrectly classifying a normal patient without cancer as part of the cancer class and the error of incorrectly classifying a patient having cancer as part of the normal class are not assessed.

In addition to the classification accuracy, both sensitivity and specificity, the receiver operating characteristic (ROC) curve, the F-measure rate or the F-1 score [22] will also be examined. The F-measure considers both the precision and recall when computing the score. In other words, it is a weighted average of the precision and recall.

2.7 Used Technologies in Existing Approaches

2.7.1 Dimensional Reduction

2.7.1.1 Principle Component Analysis(PCA)

Since the neural networks are sensitive to the number of inputs, as large number of inputs may lead to under-training, immensely large training time, loss of generality, inability to model ideal functional surfaces. In such a context, the solution lies in making effective dimensionality reduction techniques for reducing the number of input attributes, whereby having the least loss of information. The PCA performs the task of dimensionality reduction. It takes as its input a database that consists of a large number of attributes, and mines out the most interesting attributes or combination of attributes. The resultant attributes may be better suited for solving the concerned problem, and have a smaller dimensionality [23].

2.7.2 Support Vector Machine(SVM)

SVM is a supervised machine learning algorithm which can be used for classification or regression problems [22]. SVM uses for classification as well as pattern recognition purpose. It uses a function known as the kernel trick to transform data and then based on these transformations it discovers an optimal boundary between the possible outputs. Simply put, it does some extremely complex data transformations, then figures out how to separate data based on the labels or outputs which have defined [24].

Tuning parameters: Kernel, Regularization, Gamma and Margin.

Kernel

The learning of the hyperplane in linear SVM is done by transforming the problem using some linear algebra [25].

This is an equation that involves calculating the inner products of a new input vector (x) with all support vectors in training data. The coefficients must be estimated from the training data by the learning algorithm. Polynomial and exponential kernels calculates separation line in higher dimension. This is called kernel trick

Kernel Trick

The kernel function (Figure 2.2), represents a dot product of input data points mapped into the higher dimensional feature space by transformation. It expands the dataset and makes it to have more obvious boundaries between classes and the SVM algorithm is able to compute a much more optimal hyperplane [26].

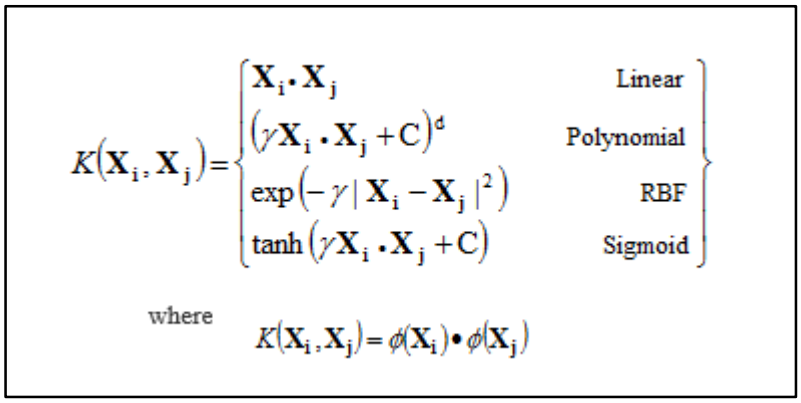

Figure 2.2:Different Kernel Functions

Gamma is an adjustable parameter of certain kernel functions. The RBF is so far the most popular choice of kernel types used in Support Vector Machines. This is mainly because of their localized and finite responses across the entire range of the real x-axis.

Gamma

The gamma parameter describes how far the impact of a single training example reaches, with low values meaning 'far' and high values meaning 'close'. In other words, with low gamma, points far away from plausible separation line are considered in calculation for the separation line. Whereas high gamma means the points close to plausible line are considered in calculation.

Regularization

The Regularization parameter (often labelled as C parameter) states the SVM optimization how much it wants to avoid misclassifying each training example.

For large values of C, the optimization will choose a smaller-margin hyperplane if that hyperplane does a better job of getting all the training points classified correctly. Conversely, a very small value of C will cause the optimizer to look for a larger-margin separating hyperplane, even if that hyperplane misclassifies more points.

Margin

A margin is a separation of line to the closest class points. A good margin is one where this separation is larger for both the classes. Figure 2.3 gives to visual example of good and bad margin. A good margin allows the points to be in their respective classes without crossing to other class.

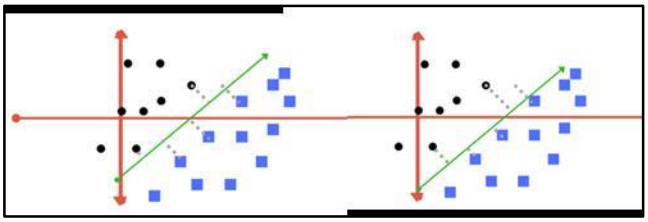

Figure 2.3: Left-Good margin; Right-Bad margin

2.7.3 K-Means Algorithm

K-means is one of the simplest unsupervised learning algorithms that solve the familiar clustering problem. It is an iterative clustering algorithm that aims to find local maxima in each iteration. In here dataset partition to k number of groups. The algorithm is included of the below steps [27]:

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.

2. Assign each object to the group that has the closest centroid.

3. When all objects have been assigned, recalculate the positions of the K centroids.

4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity. The results of the K-means clustering algorithm are [28]:

1. The centroids of the K clusters, which can be used to label new data.

2. Labels for the training data (each data point is assigned to a single cluster)

2.7.4 ROC Curve

ROC curves are frequently used to show in a graphical way the connection/trade-off between clinical sensitivity and specificity for every possible cut-off for a test or a combination of tests. ROC curves are used in clinical biochemistry to choose the most appropriate cut-off for a test [29].

The best cut-off has the highest true positive rate together with the lowest false positive rate. It shows the tradeoff between sensitivity and specificity. The area under the curve is a measure of text accuracy. To make an ROC curve have to determine true positive, true negative, false positive and false negative which is shown in Figure 2.4.

		Disease	
		+	-
Test	+	True Positive (TP)	False Positive (FP)
	-	False Negative (FN)	True Negative (TN)
		All with disease= TP + FN	All without disease= FP + TN

Figure 2.4: Comparing a method with the clinical truth

Once calculated above values then ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings (Figure 2.5).

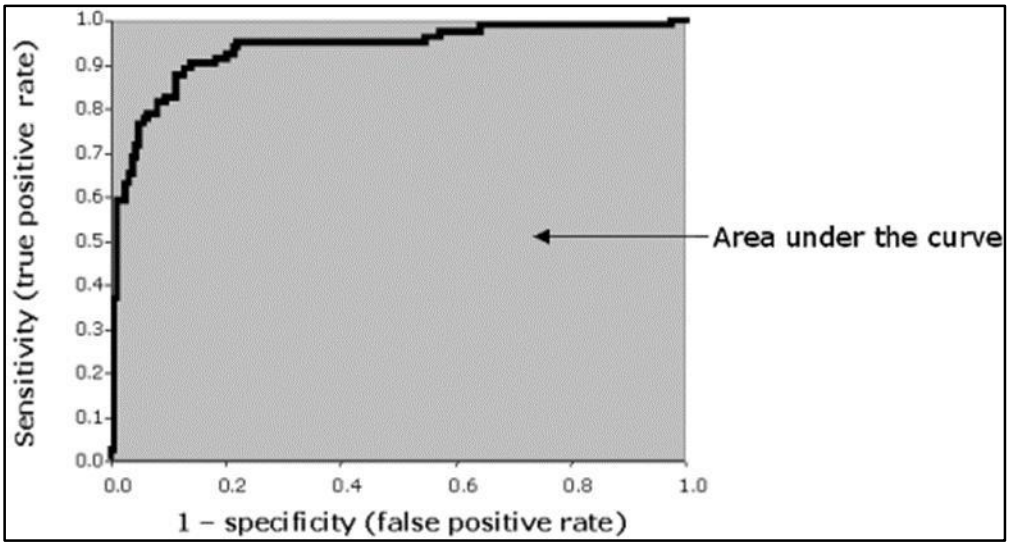

Figure 2.5: Area under ROC curve

2.7.5 Software Tools

2.7.5.1 WEKA

Weka (available on - [https://www.cs.waikato.ac.nz/ml/weka/||Weka]) is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.

All of Weka's techniques are predicated on the assumption that the data is available as one flat file or relation, where each data point is described by a fixed number of attributes [30].

2.7.5.2 R Package

R is an open source language and environment for statistical computing and graphics. It is available under the GNU General Public License, and pre-compiled binary versions are provided for different operating systems like Linux, Windows and Mac. R and its libraries implement a wide variety of statistical and graphical techniques, containing linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, and others. R is easily extensible through functions and extensions, and the R community is noted for its active contributions in terms of packages. Many of R's standard functions are written in R itself, which makes it easy for users to follow the algorithmic choices made.

CHAPTER 3

3. Research Methodology and Design

3.1 Introduction

This chapter describes the design and the methodology of the proposed solution to the research problem.

3.2 Research Methodology

The main objective of this research is to identify the correlation between clinical data and Copy Number Variation associated with hormone responsiveness genes in endometrial cancer. In order to achieve this goal exploratory type of research is carried out. There are many researches they have studied about different approaches to analyze cancer related data. They have applied different machine learning algorithms to build a proper predictions models for various areas. But there are less number of researches carried out to find the genomic alterations by studying patient clinical history. This study is focused to design a prediction model to fulfill that goal by using a proper methodology. From several paths of obtaining the goal, following design describes the most appropriate method selected by analyzing information of the literature survey.

3.3 Design

In order to design the prediction model first, the data set need to be carefully analyzed and need to preprocess the data set. Also the classes of the feature that is going to predict need to be identified accurately. After that classification will be done using a suitable algorithm and output will be analyzed to obtain the best outcome. Design process is carried out using 5 main steps which is illustrated in Figurer 3.1.

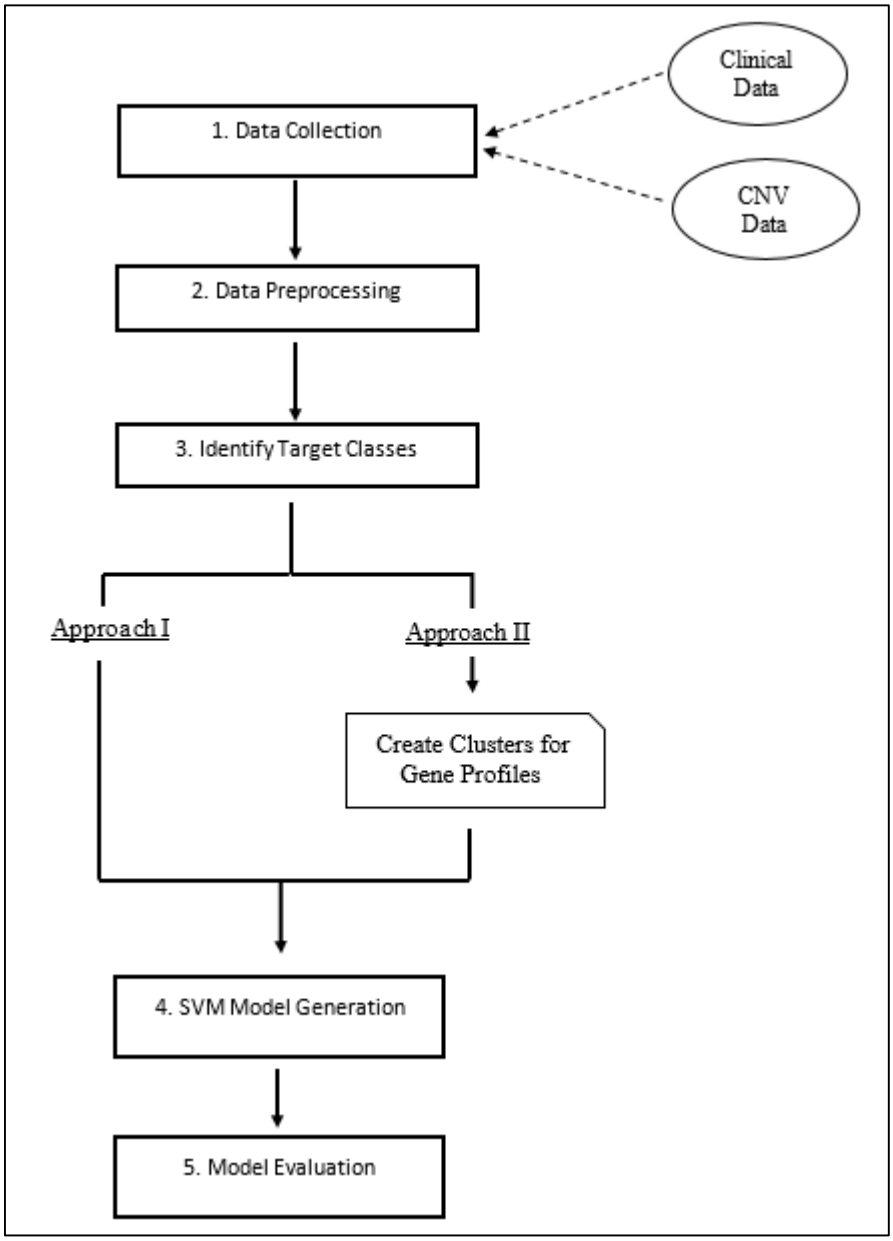

Figure 3.1: Model Design Process

In this study, we used Endometrial carcinoma data set from the cancer genome atlas(TCGA) provisional survey to generate the SVM model. Uterine corpus Endometrial Carcinoma and Uterine Carcinosarcoma are considered under endometrial cancer types. TCGA provided more than 100 of attributes for patient's clinical data and copy number variations in more than 1000 genes.

To interpret the data in a more meaningful form, it is necessary to reduce the number of variables, interpretable linear combinations of the data. In order to achieve this Principle Component Analyzing (PCA) will be used.

Preprocessed data and the previously identified hormone responsive genes are used to generate SVM model. SVM model was generated using two approaches.

1. Generates SVM model with clinical data and single gene

2. Generates SVM model with clinical data and genes profiles

Figure 3.2 shows process of Approach 1 in which the SVM model was built using preprocessed data. Validate using 10-fold cross validation. Output classes are used to make prediction.

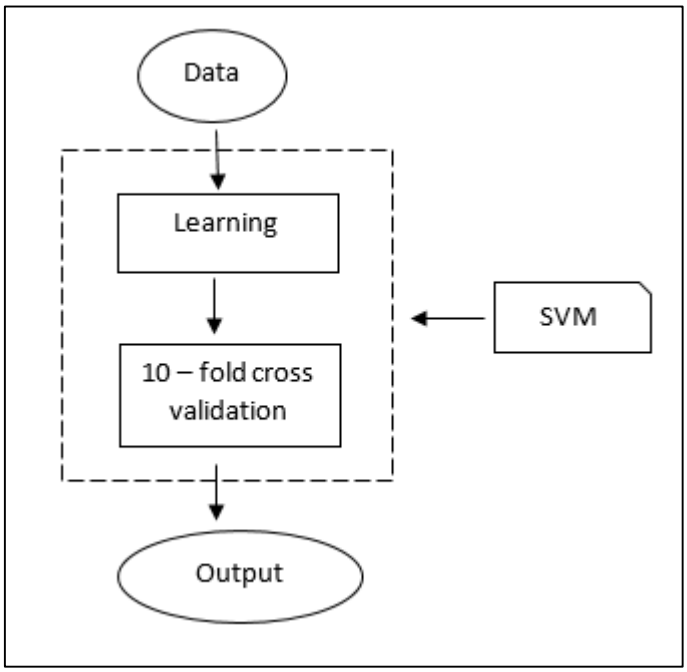

Figure 3.2: Data modeling in Approach 1

Approach 2 is extended version of approach 1. Genes are clustered before applying classification and then rest of the process is same as the approach 1. Figure 3.3 shows the process of approach 2.

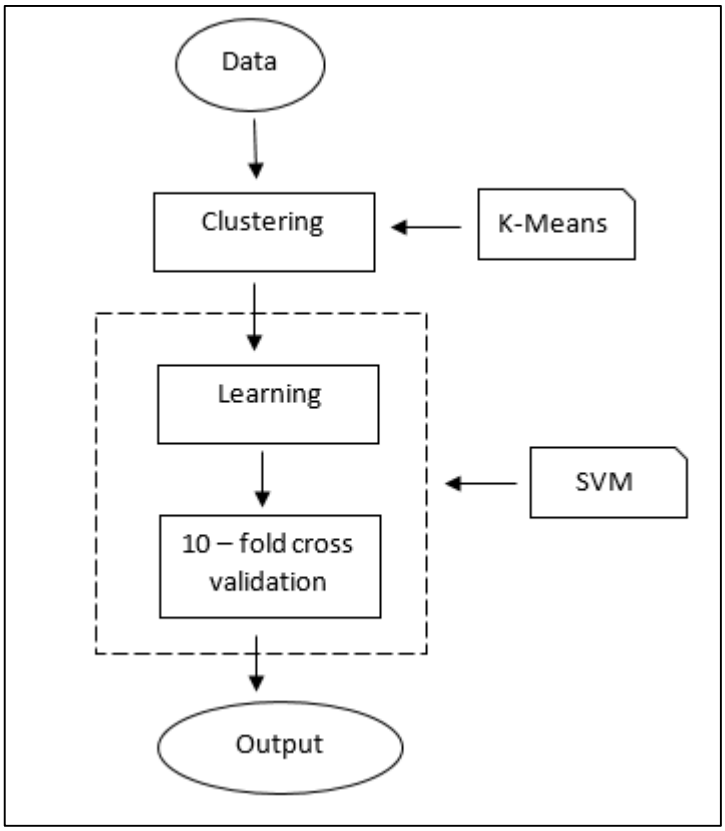

Figure 3.3: Data modeling in Approach 2

3.4 Methodology

3.4.1 Data Collection

The study is supposed to analyze tumor data of 605 endometrial cancer patients which is provided by cBio portal. Among them, 548 patient records belong to Uterine Corpus Endometrial Carcinoma and 58 of them belong to Uterine Carcinomasarcoma tumor (Figure 3.4). In this study following two types of data are mainly analyzed,

1. Patient clinical Data
2. Copy Number Variations Data

Figure 3.4: Selection of two types Endometrial Cancer

Each patient has their own clinical history including data from pre and post stages of the cancer. Each sample is with more than 100 features and it was necessary to find optimal feature set to build the model.

In addition to clinical data, Copy Number Variations data also obtained from the cBio portal (Figure 3.5). Among those, extracted the most important genes which are associated with hormone responsiveness by studying previous studies and statistics available in cBio portal.

Figure 3.5: Select CNV Data

3.4.2 Data Preprocessing

3.4.2.1 Feature Set selection

It is not always feasible to apply classification algorithms directly on dataset. First the data has to be pre-processed. Pre-processing may also involve dimensionality reduction. For High Dimensional datasets first apply Feature Selection Algorithm. The study selected 21 simple

variables commonly associated with the endometrial: age, height, weight, survival status, race category, tumor invasion, neoplasm status, main type, sub type, ethnicity category, prior cancer diagnoses, survival month, menopause status, aortic count, aortic POS by HE, aortic POS by IHC, aortic POS total, pelvic count, pelvic POS by HE, pelvic POS by IHC, pelvic POS total (Table 3.1).

Table 3.1: Description of selected variables

Variable Name	Type	Classes
Age	Numeric	
Height	Numeric	
Weight	Numeric	
Survival status	Nominal	DEAD, ALIVE
Race category	Nominal	WHITE BLACK OR AFRICAN AMERICAN, AMERICAN INDIAN OR ALASKA NATIVE, NATIVE HAWAIIAN OR OTHER PACIFIC ISLANDER and NA for missing value
Tumor invasion	Numeric	
Neoplasm status	Nominal	WITH TUMOR, TUMOR FREE
Main type	Numeric	1,2
Sub Type	Numeric	1,2
Ethnicity category	Nominal	0, -1
Prior cancer diagnoses	Nominal	YES,NO
Survival month	Numeric	
Menopause status	Nominal	POST, PERI, IND, PRE and NA for missing value
Aortic count	Numeric	
Aortic POS by HE	Numeric	
Aortic POS by IHC	Numeric	
Aortic POS total	Numeric	
Pelvic count	Numeric	
Pelvic POS by HE	Numeric	
Pelvic POS by IHC	Numeric	
Pelvic POS total	Numeric	

Feature selection was performed according to Principle Component Analysis(PCA) approach. The significance of the automatically selected set of variables was further manually evaluated by fine tuning parameters. The variables included in the final selection were those with the best discriminative performance.

For the missing values, when it be numeric then replace missing values by mean of attribute values. If the attribute is nominal then created new class label as 'NA' and replaced it.

3.4.2.2 Gene Selection of Hormone Responsiveness Genes

The next step is to explore set of hormone responsiveness genes. First the genes are identified by studying previous researches done on hormone responsiveness in endometrial cancer. Then found most significant gene by observing particular genes with cBio portal statistics.

3.4.2.3 Target Classes

According to prior researches, some genes including TXNIP CDC6 KRAS ESR1 ESRRG are recognized as important in regulating hormone responsiveness in human body They shown with maximum copy number variations (Figure 3.6).

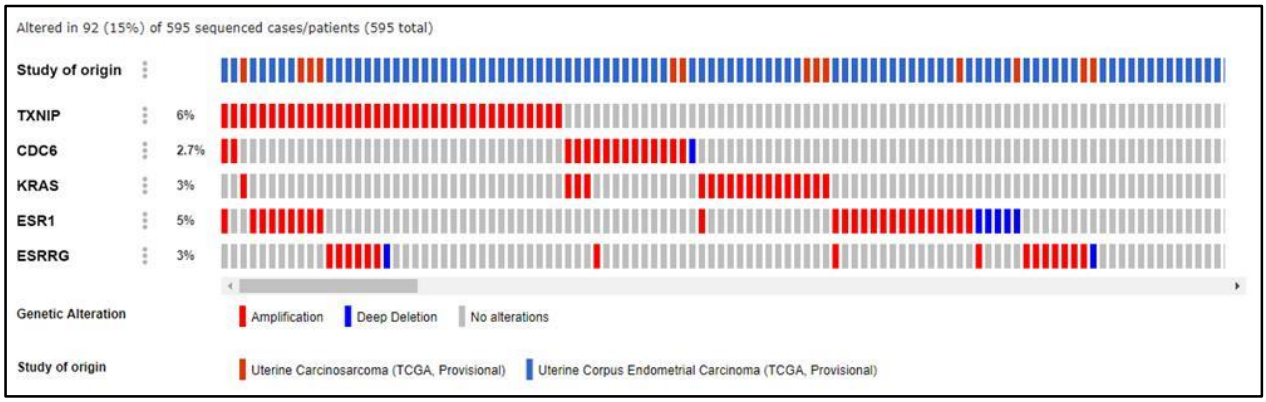

Figure 3.66:Copy Number Variations among 595 patients

3.4.3 Identify Target Classes

SVM Model is built in two approaches.

3.4.3.1 Approach I

First approach is applying the classification for one gene at a time and iteratively run it for all five genes. Output classes are shown in Table 3.2 that contain five classes. Then repeat the process for TXNIP, CDC6, KRAS, ESR1, ESRRG genes separately with clinical data and analyze the outcome.

Table 3.2: Target Classes for Approach I

Value	Description
-2	High Deletion
-1	Deletion
0	No Alteration
1	Amplification
2	High Amplification

3.4.3.2 Approach II

The second approach is make combinations of genes as shows in Table 3.3 and build SVM model. We created 3 types of combinations and used K-means to cluster and generate target classes before applying SVM.

Table 3.3: Summary of target classes in approach II

	No of Genes	Genes Description	No of Clusters
Cluster scenario 1	3	TXNIP KRAS ESR1	5
Cluster scenario 2	4	TXNIP KRAS ESR1 ESRRG	7
Cluster scenario 3	5	TXNIP KRAS ESR1 ESRRG CDC6	10

Genes can be clustered into five classes when it is considered based on the Copy Number Variation. According to Table 3.3 we reduced the number of classes from 5 to 3. If the genes altered by amplification, then it holds 1 and 2 for low amplified and high amplified respectively. In here for both phenomena indicated as 1. Similarly, if genes altered by the deletion it contains -1 and -2 values for low and high deletion. Here it is represented as -1 for both. If gene has no alterations, then value is 0. The 1st scenario has 27 patterns where number of genes is 3 and each gene contain 3 values (-1,0,1). from these 27 patters we build 5 clusters by grouping similar patterns to same cluster and different patterns to different another cluster. The next scenario using 4 genes we got 81 patterns and group all the patterns in to 7 classes. There are 243 patterns found on the last scenario when no of genes is 5. Here we create 10 clusters and grouping all 243 patterns. Finally, the outcome of these clusters is taken as the classes and SVM algorithm is applied for the classification.

3.4.4 Model Generation

3.4.4.1 Clustering

The First phase of the model is build clusters for selected genes profiles. K-means algorithm used for this. K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (5). The goal of using this algorithm is to find similar patterns (gene profiles) from the selected genes and group them as one cluster. The number of groups represented by the variable K. The algorithm works iteratively to assign each instances to one of K groups based on the features that are provided. This phase applies on the 2nd approach where genes are used as combinations.

Each of the instances merged with the closest cluster center. The closeness is determined by measuring the relative squared Euclidean distance between the instance and each cluster center. The cluster center is then updated so that each of its attribute values is the sum of the weight adjusted attribute values of the cluster center and the instance, so if the center had a weight of three and the instance had a weight of one then the resulting attribute value would be three quarters the value of the center plus one quarter the value of the instance. The weight of the instance is added to the weight of the center to create the new center weight.

When all instances have been merged into the clusters the set of clusters is passed to the classifier to build the model.

3.4.4.2 Classification

Second Phase is classification and Support Vector Machine is used to classify generated data set. The SVM algorithm performs a classification by constructing a multidimensional hyperplane that optimally discriminates between two classes by maximizing the margin between two data clusters. This algorithm achieves high discriminative power by using special nonlinear functions called kernels to transform the input space into a multidimensional space [31].

Two key parameters for the kernels, C and gamma, need to be pre-selected to generate an optimal SVM model. Parameter C controls over-fitting of the model by specifying tolerance for misclassification. Parameter gamma controls the degree of nonlinearity of the model.

In here we used WEKA [30], an open source data mining software, to generate the SVM models and Clusters. Simple K-mean which is available in WEKA, used for cluster genes. This process is carried out for 3 Scenarios. First Scenario is built 5 clusters (5 classes) for 3 genes. The second Scenario is added another gene for the previous gene set and create 7 clusters (7 classes). Final and third Scenario is added another gene to immediate previous gene set which is obtained in the second Scenario and build 10 clusters (10 classes). Then combine each outcome with the clinical data and construct SVM Models for each 3 cases.

The SMO (Sequential Minimal Optimization) is the available function in WEKA to find the optimal parameters for penalty parameter C and gamma under cross-validation. Different kernel functions, including linear, polynomial, sigmoid, and radial basis functions (RBF), were tested and selected for the models based on the performance.

Multiple logistic regression modeling (MLR) was performed using the same selected risk variables or features and case status (as specified previously and in Table 3.1) as the outcome variable.

3.4.5 Model Evaluation

Test data sets were used to assess the performance of the models. Validation using the test data sets avoided potential bias of the performance estimate due to over-fitting of the model to training data sets.

The first approach, variable values be normalized to 5 classes -2, -1, 0, 1, 2 and data points are lie between -2 to 2 range.

In the second approach we need to define clusters based on the output for gene combinations. Once the k – mean algorithm is applied on the patient CNV data, it gives different clusters based on the attributes values. The range of the spread of all data points are -1 to 1. The output is telling how each cluster comes together, with a "1" and "-1" meaning everyone in that cluster shares the same value of one, and a "0" meaning everyone in that cluster has a value of zero for that attribute. Decimal Numbers are the average value of everyone in the cluster.

3.4.5.1 10-fold cross-validation in the training data set

To evaluate the robustness of the estimates from the SVM models, a 10-fold cross-validation was performed in the training data set. The training data set was partitioned into 10 equal-size subsets. Each subset is used as a test data set for a model trained on all cases and an equal number of non-cases randomly selected from the 9 remaining data subsets. This cross-validation process is repeated 10 times, allowing each subset to serve once as the test data set. To generate summary performance estimates, we averaged the area under the curve (AUC) of the receiver operating characteristic (ROC) curve and other statistics (sensitivity, specificity, positive predictive value [PPV], negative predictive value [NPV]) of the cross-validations [32].

3.4.5.2 Statistics for performance evaluation

ROC curves were generated based on the predicted outcome and true outcome. The AUCs for the test data sets were calculated and used to compare the discriminative powers of the models. We used Delong's method [32] to calculate P-values to compare the AUCs based on results of the SVM models. Sensitivity, specificity, PPV, and NPV were calculated based on the following formulas when the cutoff value was set to default value in the SVM model (Figure 3.7).

$$Sensitivity = \frac{TP}{TP + FP} \qquad Specificity = \frac{TN}{TN + FN}$$

$$PPV = \frac{TP}{TP + FP} \qquad PPV = \frac{TP}{TP + FP}$$

Figure 3.7: Equations for finding ROC

where TP, FP, TN, and FN represent the number of true positives, false positives, true negatives, and false negatives, respectively.

CHAPTER 4

4. Results and Discussion

This chapter describes the results of applying SVM model to manipulate data sets. It shows how accurate the model is and how the accuracy changes based on different input parameters. A comparison of the two approaches used to develop the SVM model is also discussed. It summarizes the whole study and discusses the conclusions made through the study and finally shows possible directions the study can take in the future.

4.1 Approach I

We have a single csv file which contains 21 clinical attributes known as input parameters and CNV variations for one gene. We created five files for TXNIP CDC6 KRAS ESR1 ESRRG genes and converted these files to ARFF format using an online tool (available on http://ikuz.eu/csv2arff/). Then we generated SVM models by generated ARFF files.

4.1.1 SVM Model - Clinical Data with TXNIP gene

When clinical data and TXNIP gene file is passed to SVM model and target classes are -1,0,1,2. Kernel function is polynomial kernel function [33]: K (x, y) = <x, y>^p or K (x, y) = (<x, y>+1) ^p, c = 1 and exponent = 1.0. The output of the model is,

Correctly Classified Instances	172 (60.5634 %)
Incorrectly Classified Instances	112 (39.4366 %)
Kappa statistic	0.1913
Mean absolute error	0.2896
Root mean squared error	0.3699
Relative absolute error	104.4859 %
Root relative squared error	99.6094 %
Total Number of Instances	284

From the 284 instances, 172 instances are classified correctly and 112 instances goes to incorrect class. The root mean squad error is 0.37. Accuracy of the model is 60.56%. Table 4.1 indicates the performance of the SVM Model.

Table 4.1: Performance with the TXNIP gene

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.200	0.004	0.500	0.200	0.286	0.309	0.572	0.114	-1
0.854	0.724	0.593	0.854	0.700	0.159	0.568	0.589	0
0.245	0.169	0.464	0.245	0.321	0.093	0.537	0.395	1
0.000	0.000		0.000			0.396	0.048	2

4.1.2 SVM Model - Clinical Data with ESR1 gene

When clinical data and ESR1 gene file is passed to SVM model and target classes are -2, -1,0,1,2. Kernel function is polynomial kernel function: K (x, y) = <x, y>^p or K (x, y) = (<x, y>+1) ^p, c = 1 and exponent = 1.0. The output of the model is,

Correctly Classified Instances	212 (74.6479 %)
Incorrectly Classified Instances	72 (25.3521 %)
Kappa statistic	0.3014
Mean absolute error	0.2575
Root mean squared error	0.3426
Relative absolute error	146.7778 %
Root relative squared error	116.6069 %
Total Number of Instances	284

From the 284 instances, 212 instances are classified correctly and 72 instances goes to incorrect class. The root mean squad error is 0.34. Accuracy of the model is 74.64%. Table 4.2 indicates the performance of the SVM Model.

Table 4.2: Performance with the ESR1 gene

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.000	0.004	0.000	0.000	0.000	-0.004	0.495	0.004	-2
0.000	0.004	0.000	0.000	0.000	-0.016	0.429	0.064	-1
0.923	0.592	0.810	0.923	0.863	0.394	0.665	0.804	0
0.435	0.105	0.444	0.923	0.440	0.333	0.623	0.280	1
0.000	0.000		0.000			0.551	0.037	2

4.1.3 SVM Model - Clinical Data with ESRRG gene

When clinical data and ESRRG gene file is passed to SVM model and target classes are -2, -1,0,1,2. Kernel function is polynomial kernel function: K (x, y) = <x, y>^p or K (x, y) = (<x, y>+1) ^, c = 1 and exponent = 1.0. The output of the model is,

Correctly Classified Instances	169 (59.507 %)
Incorrectly Classified Instances	115 (40.493 %)
Kappa statistic	0.2049
Mean absolute error	0.2593
Root mean squared error	0.3454
Relative absolute error	114.4634 %
Root relative squared error	102.9957 %
Total Number of Instances	284

From the 284 instances, 169 instances are classified correctly and 115 instances goes to incorrect class. The root mean squad error is 0.34. Accuracy of the model is 59.50%. Table 4.3 indicates the performance of the SVM Model.

Table 4.3: Performance with the ESRRG gene

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.000	0.000		0.000			0.480	0.007	-2
0.000	0.007	0.000	0.000	0.000	-0.017	0.400	0.033	-1
0.868	0.729	0.551	0.868	0.674	0.174	0.570	0.545	0
0.258	0.146	0.564	0.258	0.354	0.140	0.559	0.461	1
0.000	0.000		0.000			0.471	0.024	2

4.1.4 SVM Model - Clinical Data with KRAS gene

When Clinical data with KRAS gene file is passed to SVM model and target classes are -1,0,1,2. Kernel function is set to polynomial kernel function: K (x, y) = <x, y>^p or K (x, y) = (<x, y>+1) ^p, c = 1 and exponent = 1.0. The output of the model is,

Correctly Classified Instances	200 (70.4225 %)
Incorrectly Classified Instances	84 (29.5775 %)
Kappa statistic	0.1811
Mean absolute error	0.2887
Root mean squared error	0.3685
Relative absolute error	129.1191 %
Root relative squared error	110.7842 %
Total Number of Instances	284

From the 284 instances, 200 instances are classified correctly and 84 instances goes to incorrect class. The root mean squad error is 0.36. Accuracy of the model is 70.42%. Table 4.4 indicates the performance of the SVM Model.

Table 4.4: Performance with the KRAS gene

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.000	0.004	0.000	0.000	0.000	-0.015	0.454	0.059	-1
0.912	0.684	0.776	0.912	0.839	0.286	0.615	0.771	0
0.260	0.124	0.310	0.260	0.283	0.146	0.551	0.207	1
0.000	0.000		0.000			0.410	0.036	2

4.1.5 SVM Model - Clinical Data with CDC6 gene

When clinical data and CDC6 gene file is passed to SVM model and target classes are -2, -1,0,1,2. Kernel function is set to polynomial kernel function: K (x, y) = <x, y>^p or K (x, y) = (<x, y>+1) ^p, c = 1 and exponent = 1.0. The output of the model is,

Correctly Classified Instances	218 (76.7606 %)
Incorrectly Classified Instances	66 (23.2394 %)
Kappa statistic	0.423
Mean absolute error	0.2535
Root mean squared error	0.3369
Relative absolute error	135.0608 %
Root relative squared error	110.6422 %
Total Number of Instances	284

From the 284 instances, 218 instances are classified correctly and 66 instances goes to incorrect class. The root mean squad error is 0.33. Accuracy of the model is 76.76%. Table 4.5 indicates the performance of the SVM Model.

Table 4.5: Performance with the CDC6 gene

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.000	0.000		0.000			0.959	0.042	-2
0.471	0.090	0.533	0.471	0.500	0.400	0.697	0.348	-1
0.940	0.476	0.825	0.940	0.879	0.532	0.732	0.818	0
0.240	0.019	0.545	0.240	0.333	0.324	0.780	0.289	1
0.000	0.000		0.000			0.831	0.145	2

Comparing 5 models for 284 instances each model has more than 60% accuracy. ESR1, TRAS and CDC6 model classified correctly 74%, 70% and 76% respectively which is considerable amount. Accuracy percentage is getting low but not least in TXNIP and ESRRG models which is 60.56% and 59.5% respectively. Figure 4.1 represent the summery of the accuracy of each model.

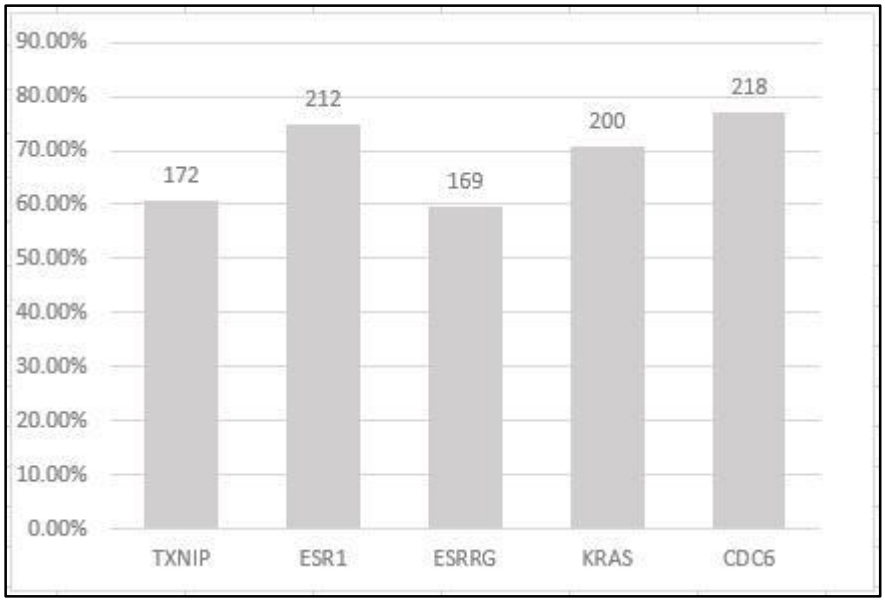

Figure 4.7:Summary of the Accuracy of each model

Normal SVM model represent binary mode which has 2 classes. But here we used multiple classes and classes has different values. The performance of the SVM model for above 5 cases shows in Table 4.6.

Table 4.6: Performance of each model

Gene	Sensitivity	Specificity	Recall	AUC
TXNIP	0.567	0.463	0.567	0.623
ESR1	0.746	0.451	0.746	0.741
ESRRG	0.549	0.432	0.549	0.690
KRAS	0.704	0.515	0.704	0.651
CDC6	0.768	0.353	0.768	0.799

4.2 Approach II

Further we have extended approach 1 by parsing combination of genes to SVM model instead of one gene. The idea behind of this is discovering genes profiles which is relate to the patient clinical data. Here we create 3 type of combinations and cluster them for create classes.

4.2.1 Scenario 1

The 1st combination is creating profile using 3 genes TXNIP, KRAS and ESR1. We have performed K-means for different k values. The optimal k is 5 and Figure 4.2 shows the output clusters.

```
kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 20.739369799764297

Initial starting points (random):

Cluster 0: 0,0,0
Cluster 1: 1,0,0
Cluster 2: 1,0,1
Cluster 3: 0,1,-1
Cluster 4: -1,1,1
```

Figure 8.2:Scenario 1- K-Mean summary

Final cluster centroids:

		Cluster#				
Attribute	Full Data	0	1	2	3	4
	(284.0)	(141.0)	(81.0)	(37.0)	(9.0)	(16.0)
TXNIP	0.412	-0.0071	1	1	0.2222	-0.125
KRAS	0.1514	0.0709	0.0617	0.2703	1	0.5625
ESR1	0.1197	-0.0284	-0.0741	1	-1	1

Figure 4.3: Cluster Centroids in Scenario 1

Cluster 0: Most of the TRNIP values are 0 in this cluster. KRAS and ESR1 attributes spread between 0 to 1 and 0 to -1 respectively. So both attributes may have 0 or 1 or -1

Cluster 1: TXNIP will be 1 for all instances in this cluster. KRAS and ESR1 probably can hold 0 or 1 or -1 and KRAS may be biased to be positive value while ESR1 may be negative.

Cluster 2: Every TXNIP and ESR1 genes in this cluster shares value 1 and KRAS will be 0 or 1 or -1

Cluster 3: KRAS and ESR1 always take 1 and -1 values respectively while TXNIP is biased to be positive value which is either 0 or 1.

Cluster 4: ESR1 holds 1 and KRAS always takes either 0 or 1. TXNIP holds 0 or -1 values.

4.2.2 Scenario 2

Next combination is added another gene for the 1st combination. Now it become 4 genes and applied k means to different k values. The algorithm sends better results when k =7. Figure 4.4 shows the output clusters.

```
kMeans
======

Number of iterations: 5
Within cluster sum of squared errors: 22.751190735102227

Initial starting points (random):

Cluster 0: 0,0,0,0
Cluster 1: 1,0,0,1
Cluster 2: 1,0,1,1
Cluster 3: 0,1,-1,1
Cluster 4: -1,1,1,-1
Cluster 5: 1,1,1,-1
Cluster 6: 0,0,-1,0
```

Figure 4.4: Scenario 2-K-Means Summary

```
Final cluster centroids:
                        Cluster#
Attribute    Full Data          0          1          2          3          4          5          6
               (284.0)    (121.0)     (76.0)     (40.0)     (10.0)     (22.0)      (5.0)     (10.0)
=============================================================================================================
TXNIP             0.412          0          1      0.825        0.1    -0.0455        0.8        0.4
KRAS             0.1514    -0.0413     0.0789      0.175        0.9          1          1       -0.1
ESR1             0.1197     0.0331    -0.0395          1       -0.9     0.2727          1       -0.9
ESRRG            0.4014     0.0579     0.9605          1        0.7    -0.1364       -0.8       -0.6
```

Figure 4.5: Cluster Centroids in Scenario 2

Cluster 0: All instances hold 0 value for TXNIP attribute. KRAS, ESR1 and ESRRG attributes lie closer to 0 values. Therefore, most of the instances take 0 for KRAS, ESR1 and ESRRG.

Cluster 1: TXNIP takes 1 for all instances in this cluster. ESRRG attribute values are closer to 1. Other two attributes, KRAS and ESR1 values are close to 0 and indicate positive and negative side.

Cluster 2: Both ESR1 and ESRRG hold 1 for every instances of this cluster. TXNIP value is close to 1 and can be 0 as well. KRAS values are close to 0 and can take 1 as well.

Cluster 3: KRAS and ESR1 attributes are close to 1 and -1 respectively. ESRRG is closer to 1 and can take 0 value also. TXNIP attribute values are close to 0.

Cluster 4: TXNIP and ESRRG attributes lie between o and -1 and most of the values can be 0. Similarly, ESR1 also close to 0 but lie on positive side. For all instances take value 1 for KRAS attribute.

Cluster 5: KRAS and ESR1 hold 1 for all instances of this cluster. TXNIP attribute is close to 1 and ESRRG is close to -1. But both attributes may have 0 as well.

Cluster 6: KRAS, ESR1 and ESRRG attributes lie between negative sides of the cluster. ESR1 attribute is close to -1 and KRAS is close to 0. TXNIP attribute lie in the middle of 0 and 1. It may hold both 0 and 1 values.

4.2.3 Scenario 3

The last combination is creating a combination by including all 5 genes and apply K-means. The optimal k value is 10 for 5 genes. Output of the scenario 3 shows in Figure 4.6.

```
kMeans
======

Number of iterations: 7
Within cluster sum of squared errors: 31.36814972383936

Initial starting points (random):

Cluster 0: 0,0,0,0,-1
Cluster 1: 1,0,0,1,0
Cluster 2: 0,0,0,0,0
Cluster 3: 1,0,1,1,1
Cluster 4: 0,1,-1,1,-1
Cluster 5: -1,1,1,-1,1
Cluster 6: 1,1,1,-1,-1
Cluster 7: 0,0,0,0,1
Cluster 8: 1,0,1,1,-1
Cluster 9: 0,0,-1,0,0
```

Figure 4.6: Scenario 3-K-Means Summary

Final cluster centroids:

Attribute	Full Data (284.0)	Cluster# 0 (15.0)	1 (75.0)	2 (116.0)	3 (21.0)	4 (6.0)	5 (7.0)	6 (6.0)	7 (11.0)	8 (20.0)	9 (7.0)
TXNIP	0.412	0.0667	0.92	0.0172	0.8095	0.5	0	0.5	-0.0909	0.9	0.7143
KRAS	0.1514	0.1333	0.0667	0.0776	-0.1429	0.6667	0.8571	1	0.5455	0.55	-0.4286
ESR1	0.1197	0.1333	-0.0133	0	1	-1	1	1	-0.6364	0.9	-0.8571
ESRRG	0.4014	0.1333	1	-0.0259	1	1	-0.1429	-0.5	-0.1818	1	-0.1429
CDC6	-0.0704	-1	0.0133	0	0.5238	-0.6667	0.7143	-1	1	-0.85	-0.8571

Figure 4.7: Cluster Centroids in Scenario 3

Cluster 0: CDC6 takes -1 value for all instances of the cluster. Most of the instances take same values for KRAS, ESR1 and ESRRG which is 0. TXNIP also is closer to 0 value.

Cluster 1: ESRRG takes 1 for all instances and TXNIP is close to 1 value. KRAS and CDC6 attributes lie on the positive side and close to 0. Even ESRRG is close to 0 and may take -1 also.

Cluster 2: Both ESR1 and CDC6 have 0 for all instances. TXNIP and KRAS are close to 0 and have probability to take both 0 and 1 values.

Cluster 3: All instances in the cluster take 1 value for attributes ESR1 and ESRRG. TXNIP and CDC6 attribute values are close to 1 and may have 0 as well. KRAS attribute lies between 0 and -1 and can hold both 0 and -1 values.

Cluster 4: ESR1 and ESRRG hold -1 and 1 values respectively. Since TXNIP has 0.5 it may hold both 0 and 1 values. For most of the instances, KRAS and CDC6 hold 0 or 1 and 0 or -1 respectively.

Cluster 5: TXNIP takes 0 and ESR1 take 1 for all instances in the cluster. Both CDC6 and KRAS attributes are close to 1 and can take both 0 and 1 values. ESRRG is close to 0 and has -1 for some instances.

Cluster 6: KRAS and ESR1 share value 1 and CDC6 takes -1 for all instances in the cluster. Since TXNIP and ESRRG have average values and lie between positive and negative side in the cluster they may have 0 or 1 and 0 or -1 respectively.

Cluster 7: CDC6 has 1 for all instances of the cluster. TXNIP, ESR1 and ESRRG attributes lie on the negative side of the cluster. ESRRG and TXNIP are close to 0 while ESR1 is close to -1. But these 3 attributes may have both 0 and -1 values.

Cluster 8: All the values of TXNIP, KRAS, ESR1 and ESRRG are close to 1 and ESRRG has exactly value of 1. CDC6 attribute value is close to -1 and may take 0 as well.

Cluster 9: All the attributes lie on the negative side except for TXNIP. TXNIP is close to 1 and others are close to -1.

4.2.4 Summary of Approach II

Table 1.7: Clustering Summary of each scenario

Type	No of Attributes	No of Classes	Squared Error
Scenario 1	3	5	20.73
Scenario 2	4	7	22.75
Scenario 3	5	10	31.36

For each Scenario, when the no of attributes gets increased even when the no of clusters gets increased, the Squared Error becomes larger. To reduce the Squared Error, we should either increase the no of clusters or decrease the no of attributes. In other words, by increasing the no of gene profiles defined by the classes we can increase the accuracy but keeping an average value of classes would be important for classification. Therefore, the best practice is to create profiles by grouping 3 genes together.

The next step of the model is calculating correlation by combining each output with clinical data. For that, each output of scenario has to be combined with previously processed clinical data and apply to SVM algorithm. Section below describes the output of SVM model.

The system generates different output for different kernel functions. Then we SVM parameters C and Gamma values for fine tune the output. RBF Kernel function gives the better output and optimal C and G values for minimum error as follows.

C = 2.0

G = 0.1

Apparently, the accuracy of this model varies around 50%. As shown in Table 4.8, the model only gives over 50% of accuracy only in Scenario 1 while giving 42.95% and 44.36% respectively in other two scenarios. These results clearly show that we cannot use this approach to make a good prediction.

Table 4.8: Classification summery for 3 scenarios

	Correctly Classified		**Incorrectly Classified**	
Scenario 1	153	53.87%	131	46.12%
Scenario 2	122	42.95%	162	57.04%
Scenario 3	126	44.36	158	55.63%

4.3 Comparisons

Here we have used two approaches to generate the SVM model. When comparing results given by these two approaches, approach 1 has produced results with the high accuracy. Out of five models created by using approach 1, there has given results with more than 70% accuracy while other two has given a 60% accuracy after putting into fine tuning. Each model produces results with the error of about 0.3. By using this model to analyze clinical data, we can predict alterations of genes based on CNV. SVM model uses Polynomial kernel function to perform the above analysis.

Approach 2 only produces optimized results when it uses RBK kernel function. But we have to perform clustering before use SVM model for classification. Because in approach 2 we use a gene profile which contains more than one gene instead of a single gene. We have proved clearly that an accurate prediction can be made on gene alteration by using the approach 1. But here by using approach 2, we tried to extend our study to predict alteration in a gene profile, but the results were not as strong as approach 1.

When we use clusters with K mean algorithm, the squad error was about 20-30. When number of genes in a gene profile increases, the number of clusters needs to be increased in order

to increase the accuracy. But as the number of clusters increases, it becomes harder to clarify what these clusters actually imply. Therefore, we decided to use less no of clusters with the model. But apparently the accuracy of the results has decreased. We also limited the no of classes indicated by a gene to 3(-1, 0, 1) where each value indicates whether the gene is amplified, deleted or altered. These limitations put approach 2 at a disadvantages over approach 1.

It is possible to define whether the amount of amplification or deletion is either Low or High by using Approach 1. Therefore, when comparing results given by the two approaches, approach 1 can be used to make better prediction over the extended model. That is to say that by using the proposed SVM model, a strong and an accurate prediction can be made on possible alteration of hormone responsive gene of a patient by analyzing the clinical history.

CHAPTER 5

5. Conclusion and Future Works

5.1 Conclusion

In this study, we tested classification schemes to determine the degree of severity of identified Hormone Responsiveness Genes with clinical data in endometrial cancer. First we create model with single gene and then extended it to test with genes profiles. Both scenarios are examples of the potential use of support vector machine techniques in the classification with different Kernel functions.

To our knowledge, this is the first attempt to study SVM approach detecting the alterations of genes by using clinical data, without doing any laboratory tests. Based on these results, we are able to predict how far genes are altered based on Copy Number variations. Even it applies to only hormone responsive genes it can be applied to other genes as well in future. Normally the process of observing and determining alterations of genes is costly and time consuming. But this proposed model can be introduced as a solution for the above process.

SVM is a model-free method that provides efficient solutions to classification problems without any assumption regarding the distribution and interdependency of the data. SVM technique has the potential to perform better than traditional statistical methods, especially in situations that include multivariate risk factors with small, limited sample size, and a limited knowledge of underlying biological relationships among risk factors. Our work provides a promising proof of principle by demonstrating the predictive power of the SVM with just a small set of variables. This approach can be extended to include large data sets, including many other variables, such as expression data, SNps as data become available.

By using SVM as a classifier, it is evident that the combination of SVM with polynomial kernel functions can yield high classification accuracy compared with the other Kernel Functions. The SVM learner gives good accuracy via both 10-Fold CV validation methods. The accuracy of the learning algorithms varies depending on the change of number of clusters.

5.2 Future Work

In this study we have extended the model used to predict alteration on a single gene to make prediction on a gene profile. There we created clusters and passed them to the SVM model as classes for the classification. But the extended model could not produce strong results as expected. Hence it is important to find a better approach which can make accurate predictions on a gene profile instead of a single gene because we believe it would be of a great value in determining the best treatment for a patient.

6. References

[1] "Metastatic Cancer," National Cancer Institute, 17 February 2017. [Online]. Available: https://www.cancer.gov/types/metastatic-cancer. [Accessed 02 October 2017].

[2] A. Sen, D. Aune and L. J. Vatten, "Hypertension and the risk of endometrial cancer: a systematic review and meta-analysis of case-control and cohort studies," vol. 10, 2017.

[3] "Uterine Carcinosarcoma," The Cancer Genome Atlas, May 2017. [Online]. Available: https://cancergenome.nih.gov/cancersselected/UterineCarcinosarcoma. [Accessed 23 October 2017].

[4] "Integrative Onco Genomics," December 2014. [Online]. Available: http://www.intogen.org/search. [Accessed 4 January 2018].

[5] "Endometrial cancer (cancer of the lining of the womb) statistics," World Cancer Research Fund International, 2014. [Online]. Available: http://www.wcrf.org/int/cancer-facts-figures/data-specific-cancers/endometrial-cancer-cancer-lining-womb-statistics. [Accessed 25 August 2017].

[6] S. M. Clinic, "Endometrial cancer," Mayo Clinic, 18 March 2017. [Online]. Available: http://www.mayoclinic.org/diseases-conditions/endometrial-cancer/home/ovc-20205704. [Accessed 19 July 2017].

[7] T. Buckers, D. Bender and K. K. Leslie, "Hormones and Receptors in Endometrial Cancer," *Proceedings in Obstetrics and Gynecology*, vol. 2, 7 2011.

[8] C. A. Walsh, "Analysis of Progesterone-regulated Target Genes in Mammary Gland and Uterine Development," *Dublin Institute of Technology ARROW@DIT*, 1 2009.

[9] J. W. Jeong, H. S. Lee, K. Y. Lee, L. D. White, R. R. Broaddus, Y. Zhang, V. Woude, L. C. Giudice, S. L. Young, B. A. Lessey, S. Y. Tsai, J. P. Lydon and F. J. DeMayo, "Mig-6 modulates uterine steroid hormone responsiveness and exhibits altered expression in endometrial disease," vol. 21, pp. 8677-8682, 2009.

[10] E. J. Hong, S. H. Park, K. C. Choi, P. C. Leung and E. B. Jeung, "Identification of estrogen-regulated genes by microarray analysis of the uterus of immature rats exposed to endocrine disrupting chemicals," *National Institutes of Health,* 2006.

[11] K. Ikeda, K. Horie-Inoue and S. Inoue, "Identification of estrogen-responsive genes based on the DNA binding properties of estrogen receptors using high-throughput sequencing technology," *Acta Pharmacologica Sinica,* vol. 10, pp. 24-31, 2015.

[12] "Copy-number variation," wikipedia, 18 April 2017. [Online]. Available: https://en.wikipedia.org/wiki/Copy-number_variation. [Accessed 20 August 2017].

[13] Wikipedia, "Copy number analysis," wikipedia, 24 April 2017. [Online]. Available: https://en.wikipedia.org/wiki/Copy_number_analysis. [Accessed 20 August 2017].

[14] "Genomic tools to detect and analyze copy number variants," Illumina, 10 December 2016. [Online]. Available: https://www.illumina.com/techniques/popular-applications/genotyping/copy-number-variation-analysis.html. [Accessed 5 September 2017].

[15] U. L. "What are single nucleotide polymorphisms (SNPs)," U.S National Library of Medicine, 3 July 2017. [Online]. Available: https://ghr.nlm.nih.gov/primer/genomicresearch/snp. [Accessed 13 October 2017].

[16] E. Cerami, J. Gao, U. Dogrusoz, B. E. Gross, S. O. Sumer, B. A. Aksoy, A. Jacobsen, C. J. Byrne , M. L. Heuer , E. Larsson, Y. Antipin, B. Reva, A. P. Goldberg, C. Sander and N. Schultz, "The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data," *American Association for Cancer Research,* vol. 2, pp. 401-404, 5 2012.

[17] "ICGC Data Portal," Ontarip Institute for Cancer Research, 2017. [Online]. Available: https://dcc.icgc.org/. [Accessed 1 November 2017].

[18] J. T. Robinson, H. Thorvaldsdottir, W. Wendy, M. Guttman, E. S. Lander, G. Getz and J. P. Mesirov, "Integrative Genomics Viewer (IGV)," *Nature Biotechnology,* vol. 29, pp. 24-26, 2011.

[19] A. Bharathi and A. M. Natarajan, "Cancer Classification using Support Vector Machines and Relevance Vector Machine based on Analysis of Variance Features," *Journal of Computer Science,* vol. 7, pp. 1393-1399, 2011.

[20] T. S. Furey, "Support Vector Machine classification and validation of cancer tissue samples using microarray expression data.Bioinformatics," vol. 16, pp. 906-914, 2000.

[21] M. P. Brown, W. N. Grundy, T. S. Furey, C. W. Sugnet, N. Cristianini, D. Lin, M. J. Ares and D.Haussler, "Support Vector Machine Classification of Microarray Gene Expression Data," pp. 99-9.

[22] C. D. ArockiaVanitha, "Gene Expression Data Classification Using Support Vector Machine and Mutual Information-based Gene Selection," *Procedia Computer Science,* vol. 47, pp. 13-21, 2015.

[23] R. R. Janghel, R. Tiwari, R. Kala and A.Shukla, "Breast Cancer Data Prediction by Dimensionality Reduction Using PCA and Adaptive Neuro Evolution," *International Journal of Information Systems and Social Change,* vol. 3, pp. 1-9, 1 2012.

[24] M. Abernethy, "Classification and clustering," IBM,Developer Work, 11 May 2010. [Online]. Available: https://www.ibm.com/developerworks/library/os-weka2/. [Accessed 5 January 2018].

[25] S.Patel, "Support Vector Machine Theory," 5 2017.

[26] "Support Vector Machines (SVM) Introductory Overview," StatSoft.com, [Online]. Available: http://www.statsoft.com/Textbook/Support-Vector-Machines. [Accessed 18 January 2018].

[27] "A Tutorial on Clustering Algorithms," [Online]. Available: https://home.deib.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html. [Accessed 5 January 2018].

[28] A. Trevino, "Introduction to K-means Clustering," Oracle-DataScience.com, 6 December 2016. [Online]. Available: https://www.datascience.com/blog/k-means-clustering. [Accessed 18 January 2018].

[29] T. G. Tape and M. , "Plotting and Intrepretating an ROC Curve," [Online]. Available: http://gim.unmc.edu/dxtests/roc2.htm. [Accessed 21 November 2017].

[30] "Weka (machine learning)," Wikipedia, 8 June 2017. [Online]. Available: https://en.wikipedia.org/wiki/Weka_(machine_learning). [Accessed 19 August 2017].

[31] Y.K.Alapati, "Combining Clustering with Classification:A Technique to Improve Classification Accuracy," *International Journal of Computer Science Engineering (IJCSE),* vol. 5, 11 2016.

[32] W. YuEmail, T. Liu , R. Valdez, M. Gwinn and M. J. Khoury, "Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes," *BMC Medical Informatics and Decision Making,* vol. 10, 3 2010.

[33] J. Weston, S. Mukherjee O, C. M. Pontiltt, T. Poggiott and V. Vapnik, "Feature Selection for SVMs".

[34] A. Dutt and R. Beroukhim, "Single nucleotide polymorphism array analysis of cancer," *Current Opinion in Oncology*, vol. 19, pp. 43-49, 2007.

7. Appendix

Code Samples

Apply PCA algorithm to feature selection from the patient data set

```
#PCA testing
hd<-iris(,-5)
head(d)
pc<-princomp(d, cor=TRUE, score=TRUE)
summary(pc)

plot(pc)
plot(pc, type="1")
biplot(pc)
dim(d)
attributes(pc)
pc$loadings
```

Figure 7.1:PCA algorithm sample code

Apply Linear Regression algorithm on the data set

```
#Load Train and Test datasets
#Identify feature and response variable(s) and values must be numeric and numpy arrays

#input_variables_values_training_datasets
x_train <- read.csv("C:\\Users\\uhtharas\\Documents\\Research\\
clinicalVstype\\clinicalDataVsType_withoutNA_Random.csv", header=TRUE)

#target_variables_values_training_datasets
y_train <- "main_type"

#input_variables_values_test_datasets
x_test <- read.csv("C:\\Users\\uhtharas\\Documents\\Research\\
clinicalVstype\\clinicalDataVsType_withoutNA_Random_test.csv", header=TRUE)

x <- cbind(x_train,y_train)

# Train the model using the training sets and check score
linear <- lm(y_train ~ ., data = x)
summary(linear)
#Predict Output
predicted= predict(linear,x_test)
```

Figure 7.2: SVM sample code

Apply SVM algorithm on the data set

```
install.packages('e1071', dependencies=TRUE)


library(e1071)
#input_variables_values_training_datasets
data=read.csv("C:\\Users\\uhtharas\\Documents\\Research\\
clinicalVstype\\clinicalDataVsType_withoutNA_Random.csv",
 header=TRUE)
model <- svm(main_type ~ ., data)
print(model)
summary(model)

#target_variables_values_training_datasets
y_train <- "norm"

#input_variables_values_test_datasets
x_test <- read.csv("C:\\Users\\uhtharas\\Documents\\
Research\\clinicalVstype\\clinicalDataVsType_withoutNA_Random_test.csv",
 header=TRUE)

x <- cbind(x_train,y_train)
m <- svm(x_train, y_train)
# Fitting model
fit <-svm(y_train ~ ., data = x)
summary(fit)
#Predict Output
predicted= predict(fit,x_test)
```

Figure 9:SVM code sample

Data Sheet

id	age	height	weight	survival_status	race_cat	tumer_invation	neoplsma_status	main_type	sub_type	ethnicity_cat	prior_cancer_diagnose	survival_month	menopause_status	aortic_count	aortic_pos_by_he	aortic_pos_by_ihc	aortic_pos_total	pelvic_count	pelvic_pos_by_he	pelvic_pos_by_ihc	pelvic_pos_total	TXNIP	KRAS	ESR1	ESRRG	CDC6	schema 1	schema 2	schema 3
121	76	163	60	Dead	WHITE	82	WITH TUMOR	1	2	-1	No	46.75	POST	1	0	0	0	20	6	0	6	0	1	1	0	-1	4	4	6
155	77	152	60	Alive	WHITE	50	NA	1	2	0	No	34.63	POST	9	0	0	0	24	0	0	0	0	0	-1	-1	1	0	6	7
156	77	156	66	Alive	WHITE	50	NA	1	2	0	No	25.49	POST	3	0	0	0	12	1	0	1	1	0	0	1	-1	1	1	1
166	84	156	55	Alive	WHITE	0	NA	1	2	0	Yes	36.37	POST	5	0	0	0	11	0	0	0	0	0	0	0	0	0	0	2
240	79	147	44	Dead	WHITE	100	WITH TUMOR	1	2	-1	No	50.85	POST	0	0	0	0	0	1	0	1	0	0	-1	1	0	0	3	4
477	71	161	62	Dead	AFRICAN A	50	WITH TUMOR	1	2	0	No	22.01	POST	0	0	0	0	9	7	0	7	1	1	1	1	1	2	2	3
244	62	146	67	Dead	AFRICAN A	100	WITH TUMOR	1	2	-1	No	11.86	POST	5	0	0	2	0	1	0	1	1	2	-1	0	2	3	3	7
431	76	0	65	Alive	AFRICAN A	39.4	TUMOR FREE	1	2	0	No	79.04	POST	5	0	0	0	26	0	0	0	1	1	-1	-1	2	3	6	7
397	74	151	93	Alive	WHITE	16	TUMOR FREE	1	2	0	No	13.21	POST	8	0	0	0	37	0	0	0	0	0	0	0	-1	0	0	0
214	71	160	61	Alive	WHITE	5	TUMOR FREE	1	2	-1	No	27.83	POST	0	0	0	0	12	1	0	0	0	1	2	0	-1	4	4	6
485	83	159	81	Dead	WHITE	45	TUMOR FREE	1	2	0	Yes	10.18	POST	19	0	0	0	33	1	0	1	0	0	2	1	0	4	2	3
49	61	152	141	Dead	WHITE	39.4	WITH TUMOR	1	2	-1	No	20.24	POST	0	0	0	0	4	0	0	1	1	1	1	-1	-1	2	5	6
439	63	161	79	Alive	NA	100	TUMOR FREE	1	2	-1	No	10.91	POST	1	1	0	1	17	8	0	8	1	1	1	1	0	2	2	8
441	87	146	58	Alive	NA	40	TUMOR FREE	1	2	-1	No	37.68	POST	0	0	0	0	16	0	0	0	1	0	-1	0	-1	1	6	9
297	57	179	135	Alive	WHITE	39.4	NA	1	1	0	No	23.69	PERI	5	0	0	0	8	1	0	1	0	1	0	0	0	0	4	2
399	66	159	87	Alive	WHITE	95	TUMOR FREE	1	2	0	No	18.23	POST	23	13	0	13	39	8	0	8	0	0	0	0	-1	0	0	0
425	63	173	71	Alive	WHITE	3	TUMOR FREE	1	2	0	No	35.35	POST	9	2	0	2	10	0	0	0	0	1	1	0	1	4	4	5
170	67	161	87	Alive	AFRICAN A	39.4	NA	1	2	0	No	71.19	NA	5	0	0	3	13	1	0	9	0	0	0	0	-1	0	0	0
176	67	157	70	Alive	WHITE	39.4	TUMOR FREE	1	2	0	Yes	27.86	POST	9	0	0	0	27	1	0	0	0	2	0	-1	0	0	4	2
505	66	168	166	Dead	WHITE	33	TUMOR FREE	1	3	0	No	77.27	POST	0	0	0	0	0	1	0	1	1	1	-1	1	-1	3	3	4
510	71	160	83	Alive	AFRICAN A	75	WITH TUMOR	1	2	-1	No	58.97	POST	8	0	0	0	29	0	0	0	1	0	1	1	0	2	2	3
511	58	154	89	Alive	WHITE	13	WITH TUMOR	1	2	0	No	9.33	POST	6	0	0	0	28	0	0	0	2	-1	-1	1	1	1	1	1
514	64	173	99	Alive	WHITE	26	TUMOR FREE	1	2	0	No	30.65	POST	6	0	0	0	24	0	0	0	2	0	1	2	-1	2	2	8
85	69	159	97	Alive	WHITE	37	TUMOR FREE	1	1	-1	No	129.7	POST	0	0	0	0	0	1	0	1	1	0	0	1	0	1	1	1
86	59	165	126	Dead	AFRICAN A	0	WITH TUMOR	1	2	-1	No	21.98	POST	0	0	0	0	0	0	0	0	1	-1	0	2	-1	1	1	1
87	90	156	48	Dead	WHITE	25	TUMOR FREE	1	1	0	Yes	110.55	POST	3	0	0	0	12	0	0	0	1	0	0	0	-1	1	1	0
89	64	151	56	Alive	WHITE	96	TUMOR FREE	1	1	-1	No	86.01	POST	18	0	0	0	20	0	0	0	0	0	0	0	0	0	0	2

91	81	0	0	Dead	WHITE	91	TUMOR FREE	1	2	-1	No	63.86	POST	0	0	0	0	6	0	0	0	2	0	0	0	-1	1	1	0
92	67	161	98	Alive	WHITE	73	TUMOR FREE	1	2	-1	Yes	57.75	POST	8	1	0	1	29	1	0	1	0	0	-1	0	-1	0	6	9
93	75	153	89	Alive	WHITE	61	TUMOR FREE	1	2	-1	No	45.5	POST	7	0	0	0	24	0	0	0	2	-1	0	0	-1	1	0	9
94	66	168	79	Alive	WHITE	0	TUMOR FREE	1	2	-1	No	36.86	POST	1	0	0	0	14	0	0	0	0	0	0	0	-1	0	0	0
95	58	162	123	Dead	WHITE	23	TUMOR FREE	1	1	-1	No	23.85	PRE	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	2
131	81	152	48	Dead	WHITE	50	WITH TUMOR	1	1	0	No	13.57	POST	4	0	0	0	18	5	0	5	1	0	0	1	0	1	1	1
132	82	161	61	Alive	WHITE	50	TUMOR FREE	1	1	-1	No	33.18	POST	7	0	0	0	14	0	0	0	0	0	0	0	0	0	0	2
133	56	155	62	Alive	DIAN OR A	50	TUMOR FREE	1	1	0	No	29.66	POST	11	0	0	0	12	0	0	0	0	0	0	0	0	0	0	2
134	60	161	74	Alive	WHITE	50	TUMOR FREE	1	1	0	No	80.49	POST	6	0	0	0	15	0	0	0	1	1	0	1	0	1	1	1
135	57	166	92	Alive	WHITE	50	TUMOR FREE	1	1	0	No	88.17	POST	6	0	0	0	24	0	0	0	1	0	0	1	0	1	1	1
136	37	150	66	Alive	NA	50	TUMOR FREE	1	1	1	No	71.45	IND	7	0	0	0	13	0	0	0	1	0	0	1	0	1	1	1
137	77	163	87	Alive	DIAN OR A	50	TUMOR FREE	1	1	0	No	35.32	POST	6	0	0	0	17	0	0	0	1	0	0	1	0	1	1	1
138	53	173	122	Alive	WHITE	50	TUMOR FREE	1	1	-1	No	45.96	IND	8	0	0	0	25	0	0	0	0	0	0	0	0	0	0	2
139	63	160	88	Alive	WHITE	50	TUMOR FREE	1	1	0	No	79.4	POST	3	0	0	0	19	0	0	0	1	0	0	1	0	1	1	1
140	81	154	48	Alive	WHITE	50	TUMOR FREE	1	1	-1	No	49.84	POST	5	0	0	0	29	0	0	0	1	0	0	1	0	1	1	1
141	72	151	93	Alive	WHITE	50	TUMOR FREE	1	1	0	Yes	67.61	POST	8	0	0	0	13	0	0	0	0	-1	0	1	0	0	0	1
301	57	168	127	Alive	WHITE	40	TUMOR FREE	1	1	0	No	35.87	POST	4	0	0	0	18	0	0	0	1	0	1	1	0	2	2	3
304	69	152	72	Alive	AFRICAN A	54	WITH TUMOR	1	2	1	No	35.05	POST	5	0	0	0	12	2	0	2	0	0	1	1	-1	4	2	8
7	73	160	48	Alive	AFRICAN A	39.4	TUMOR FREE	1	2	-1	No	36.99	POST	9	0	0	0	21	0	0	0	1	0	0	0	0	1	1	2
8	79	172	82	Alive	WHITE	70	TUMOR FREE	1	1	-1	No	83.44	POST	5	0	0	0	7	0	0	0	1	0	0	1	0	1	1	1
9	67	170	90	Dead	WHITE	46	WITH TUMOR	1	1	0	No	17.84	POST	3	3	0	3	3	3	0	3	1	0	0	1	0	1	1	1
10	65	165	122	Alive	WHITE	39.4	TUMOR FREE	1	1	-1	No	32.75	POST	10	0	0	0	5	0	0	0	1	0	0	1	0	1	1	1
11	75	163	74	Dead	WHITE	39.4	TUMOR FREE	1	1	-1	No	112.45	POST	3	0	0	0	9	0	0	0	0	0	0	0	0	0	0	2
12	38	155	57	Alive	ASIAN	39.4	TUMOR FREE	1	1	0	No	105.29	PRE	5	0	0	0	3	0	0	0	0	1	0	0	0	0	4	2
14	57	163	55	Alive	WHITE	33	TUMOR FREE	1	1	0	No	74.05	POST	5	0	0	0	16	0	0	0	0	0	0	0	0	0	0	2
15	63	168	66	Alive	WHITE	19	TUMOR FREE	1	1	0	Yes	88.11	POST	1	0	0	0	6	0	0	0	1	0	0	1	0	1	1	1
16	44	163	115	Alive	WHITE	0	TUMOR FREE	1	1	0	No	65.7	POST	3	0	0	0	4	0	0	0	0	0	0	1	-1	0	0	0
17	53	160	97	Alive	WHITE	49	TUMOR FREE	1	1	0	Yes	60.48	POST	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	2
18	65	164	68	Alive	WHITE	56	TUMOR FREE	1	1	0	No	60.48	POST	0	0	0	0	5	0	0	0	2	0	-1	1	-1	1	1	4
19	58	160	60	Alive	ASIAN	39.4	TUMOR FREE	1	1	0	No	46.52	POST	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	2
22	58	160	95	Alive	WHITE	18	TUMOR FREE	1	1	-1	No	28.25	POST	0	0	0	0	10	0	0	0	1	0	0	1	0	1	1	1
23	67	158	63	Alive	WHITE	39.4	TUMOR FREE	1	1	0	No	53.98	POST	0	0	0	0	21	0	0	0	1	0	0	1	0	1	1	1
24	46	142	49	Alive	ASIAN	7	TUMOR FREE	1	1	0	No	23.19	POST	5	0	0	0	4	0	0	0	0	0	0	0	0	0	0	2
25	53	164	108	Alive	WHITE	10	TUMOR FREE	1	1	-1	No	68.59	POST	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1

88	64	167	75	Dead	WHITE	100	WITH TUMOR	1	1	0	No	23.29	POST	4	0	0	0	22	1	0	1	1	0	0	1	0	1	1	1
97	54	161	74	Alive	ASIAN	0	TUMOR FREE	1	1	0	No	20.89	PRE	0	0	0	0	4	0	0	0	1	1	0	1	0	1	1	1
98	70	0	61	Dead	AFRICAN A	87	WITH TUMOR	1	2	-1	No	48.75	POST	6	1	0	1	6	0	0	0	-1	0	1	1	-1	4	0	0
99	71	152	60	Alive	WHITE	0	TUMOR FREE	1	2	-1	No	125.33	POST	2	0	0	0	12	0	0	0	0	1	-1	2	-1	3	3	4
100	63	158	68	Alive	WHITE	87	TUMOR FREE	1	1	-1	No	117.9	POST	12	0	0	0	20	0	0	0	0	0	0	1	0	0	0	1
103	54	163	76	Alive	WHITE	18	TUMOR FREE	1	1	-1	No	66.2	POST	2	0	0	0	3	0	0	0	0	0	0	0	0	0	0	2
104	60	164	83	Dead	WHITE	0	WITH TUMOR	1	3	0	No	36.33	POST	14	0	0	0	17	0	0	0	1	0	0	1	0	1	1	1
105	67	161	86	Alive	AFRICAN A	83	TUMOR FREE	1	2	-1	No	97.63	POST	3	2	0	2	26	1	0	1	2	-1	1	1	0	2	2	3
106	42	152	80	Alive	WHITE	0	TUMOR FREE	1	1	-1	No	46.68	PRE	6	0	0	0	14	0	0	0	0	0	0	0	0	0	0	2
109	56	166	119	Alive	AFRICAN A	0	TUMOR FREE	1	1	-1	No	82.46	POST	0	0	0	0	36	0	0	0	0	0	0	0	0	0	0	2
110	45	160	67	Dead	WHITE	0	WITH TUMOR	1	1	0	No	33.38	PRE	6	0	0	0	15	0	0	0	2	-1	1	2	0	2	2	3
114	57	165	91	Alive	WHITE	38	TUMOR FREE	1	1	-1	No	49.18	POST	0	0	0	0	16	0	0	0	0	0	0	0	0	0	0	2
115	39	152	76	Alive	WHITE	36	TUMOR FREE	1	1	-1	No	25.07	PRE	0	0	0	0	16	0	0	0	0	0	0	0	0	0	0	2
149	67	163	61	Alive	WHITE	50	TUMOR FREE	1	2	0	No	37.75	POST	9	0	0	0	24	0	0	0	1	1	-1	1	0	3	3	4
152	80	160	57	Alive	WHITE	50	TUMOR FREE	1	1	0	No	77.33	POST	7	0	0	0	15	0	0	0	1	1	0	1	0	1	1	1
199	73	161	105	Alive	WHITE	71	TUMOR FREE	1	1	-1	No	56.87	POST	6	0	0	0	32	0	0	0	1	0	0	1	0	1	1	1
206	60	162	87	Alive	WHITE	61	TUMOR FREE	1	1	-1	No	71.39	POST	8	0	0	0	19	1	0	1	0	0	0	0	0	0	0	2
207	48	168	140	Alive	WHITE	0	TUMOR FREE	1	1	-1	No	48.55	IND	0	0	0	0	15	1	0	0	1	0	0	1	0	1	1	1
208	69	174	117	Alive	WHITE	45	TUMOR FREE	1	1	-1	No	36.47	POST	0	0	0	0	22	1	0	0	0	0	0	0	0	0	0	2
210	62	164	88	Dead	AFRICAN A	10	WITH TUMOR	1	1	-1	No	27.27	POST	11	0	0	0	27	1	0	0	0	2	2	-1	-1	4	5	6
211	51	162	75	Alive	NA	87	TUMOR FREE	1	1	-1	No	57.72	POST	3	0	0	0	23	1	0	0	1	0	0	1	0	1	1	1
212	58	167	132	Alive	WHITE	7	TUMOR FREE	1	1	-1	No	64.16	POST	0	0	0	0	0	1	0	1	1	0	0	1	0	1	1	1
257	47	157	141	Alive	WHITE	5	TUMOR FREE	1	1	0	No	18.59	IND	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	2
258	57	170	96	Alive	WHITE	90	TUMOR FREE	1	1	0	No	20.17	POST	7	0	0	0	15	0	0	0	1	0	0	1	0	1	1	1
264	60	160	102	Alive	WHITE	75	TUMOR FREE	1	1	0	No	63.63	POST	5	0	0	0	16	1	0	1	0	0	0	0	0	0	0	2
268	74	155	135	Alive	WHITE	0	TUMOR FREE	1	1	0	No	41.49	POST	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	2
271	83	165	101	Dead	WHITE	13	TUMOR FREE	1	1	0	No	23.46	POST	2	0	0	0	13	0	0	0	0	0	0	0	0	0	0	2
273	63	160	83	Alive	WHITE	35.7	TUMOR FREE	1	1	0	No	43	POST	5	0	0	0	16	1	0	1	1	0	0	1	0	1	1	1
26	64	157	68	Alive	WHITE	82	TUMOR FREE	1	2	1	No	37.88	POST	3	0	0	0	17	1	0	1	0	0	-1	0	0	0	6	2
27	55	165	136	Alive	WHITE	40	TUMOR FREE	1	1	0	No	17.58	POST	8	0	0	0	8	0	0	0	0	1	0	0	-1	0	4	0
28	81	160	62	Alive	WHITE	57	TUMOR FREE	1	1	0	No	19.58	POST	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	2
29	51	163	145	Alive	WHITE	49	TUMOR FREE	1	1	0	No	24.61	POST	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	2
30	68	155	113	Alive	WHITE	75	TUMOR FREE	1	1	0	No	29.04	POST	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	2
32	74	163	58	Alive	WHITE	39.4	TUMOR FREE	1	1	0	No	42.31	POST	0	0	0	0	14	0	0	0	0	0	0	0	0	0	0	2

33	62	158	89	Alive	NA	36	TUMOR FREE	1	1	1	No	15.93	POST	12	0	0	0	12	0	0	0	0	0	0	0	0	0	0	2
101	57	160	75	Alive	WHITE	5	TUMOR FREE	1	1	-1	No	110.28	POST	5	0	0	0	19	0	0	0	1	0	0	1	0	1	1	1
107	56	161	72	Alive	WHITE	8	WITH TUMOR	1	1	-1	No	83.9	POST	8	0	0	0	29	0	0	0	1	0	0	1	0	1	1	1
111	76	167	60	Alive	WHITE	12	TUMOR FREE	1	1	-1	No	61.6	POST	10	0	0	0	17	0	0	0	0	0	0	0	0	0	0	2
112	59	163	95	Alive	WHITE	5	TUMOR FREE	1	1	-1	No	80.72	POST	6	0	0	0	33	0	0	0	0	0	0	0	0	0	0	2
148	79	158	76	Alive	AFRICAN A	50	TUMOR FREE	1	2	0	No	35.71	POST	6	1	0	1	7	1	0	1	0	0	0	0	0	0	0	2
198	84	154	58	Alive	AFRICAN A	0	TUMOR FREE	1	2	-1	Yes	80.32	POST	0	0	0	0	0	0	0	0	2	1	2	-1	2	2	5	5
200	54	170	110	Alive	WHITE	11	TUMOR FREE	1	1	-1	No	86.3	POST	0	0	0	0	28	0	0	0	0	0	0	0	0	0	0	2
203	63	166	94	Alive	AFRICAN A	26	TUMOR FREE	1	1	-1	No	85.94	POST	2	0	0	0	16	0	0	0	0	0	0	1	0	0	0	1
205	50	168	67	Alive	WHITE	100	TUMOR FREE	1	1	-1	No	69.22	POST	0	0	0	0	17	1	0	2	0	0	0	0	0	0	0	2
209	54	159	84	Alive	WHITE	0	TUMOR FREE	1	1	-1	No	72.57	IND	0	0	0	0	25	1	0	0	0	0	0	0	0	0	0	2
213	64	168	61	Alive	WHITE	0	TUMOR FREE	1	1	-1	No	63.5	POST	0	0	0	0	6	1	0	0	1	0	0	1	0	1	1	1
215	88	160	65	Dead	WHITE	87	TUMOR FREE	1	1	-1	No	13.44	POST	0	0	0	0	0	1	0	1	1	0	0	1	0	1	1	1
259	66	165	117	Alive	WHITE	6	TUMOR FREE	1	1	0	No	19.32	POST	0	0	0	0	1	0	0	0	1	0	0	1	0	1	1	1
260	62	168	76	Alive	WHITE	59	TUMOR FREE	1	1	0	No	20.93	POST	1	0	0	0	8	0	0	0	1	1	0	1	1	1	1	1
261	74	160	75	Alive	WHITE	53	TUMOR FREE	1	1	0	No	35.18	POST	5	0	0	0	12	0	0	0	1	0	0	1	0	1	1	1
262	60	178	97	Alive	WHITE	5	TUMOR FREE	1	1	0	No	71.19	POST	3	0	0	0	14	0	0	0	1	0	0	1	1	1	1	1
263	73	152	55	Dead	WHITE	100	TUMOR FREE	1	2	0	Yes	22.04	POST	5	3	0	3	16	5	0	5	1	2	0	1	2	1	1	1
265	50	160	209	Alive	WHITE	50	TUMOR FREE	1	1	0	Yes	66.36	IND	0	0	0	0	0	1	0	1	0	1	1	1	-1	4	2	8
266	73	160	93	Alive	WHITE	75	TUMOR FREE	1	1	0	No	74.57	POST	4	0	0	0	5	0	0	0	0	0	0	0	0	0	0	2
269	73	160	61	Alive	WHITE	36	TUMOR FREE	1	1	0	No	48.52	POST	2	0	0	0	5	0	0	0	0	1	0	0	0	0	4	2
277	78	152	68	Alive	WHITE	76.9	TUMOR FREE	1	1	0	No	20.27	POST	1	0	0	0	8	0	0	0	1	0	0	1	0	1	1	1
278	68	170	118	Alive	WHITE	96	TUMOR FREE	1	1	0	No	15.41	POST	0	0	0	0	9	0	0	0	1	0	0	1	0	1	1	1
303	60	154	111	Alive	WHITE	76	TUMOR FREE	1	1	0	No	35.87	POST	8	2	0	2	13	0	0	0	0	0	0	0	0	0	0	2
314	58	152	90	Dead	OR OTHER	88	WITH TUMOR	1	1	1	No	24.31	POST	1	0	0	1	14	0	0	2	0	0	0	0	0	0	0	2
315	69	147	68	Alive	ASIAN	9	TUMOR FREE	1	1	0	No	85.48	POST	3	0	0	0	20	0	0	0	0	0	0	0	0	0	0	2
316	65	155	85	Alive	ASIAN	3	TUMOR FREE	1	1	0	No	78.15	POST	0	0	0	0	13	0	0	0	0	0	0	0	0	0	0	2
317	35	168	91	Dead	ASIAN	100	WITH TUMOR	1	1	0	No	4.8	PRE	0	0	0	0	6	0	4	0	0	1	0	0	0	0	4	2
318	60	163	64	Alive	WHITE	40	TUMOR FREE	1	1	0	No	77.89	POST	1	0	0	0	7	0	0	0	0	0	0	0	0	0	0	2
319	64	163	59	Alive	WHITE	92	TUMOR FREE	1	1	0	No	61.83	POST	4	0	0	0	14	0	0	0	1	1	1	1	0	2	2	8
320	59	152	88	Alive	ASIAN	58	TUMOR FREE	1	1	0	No	67.94	POST	9	0	0	0	34	0	0	0	0	0	0	0	0	0	0	2
322	63	165	66	Dead	WHITE	51	TUMOR FREE	1	1	0	No	30.72	POST	14	0	0	0	13	0	0	0	1	0	0	1	0	1	1	1
323	55	170	82	Alive	WHITE	100	TUMOR FREE	1	1	0	No	97.34	POST	1	0	0	0	4	0	0	3	0	0	0	0	0	0	0	2
326	65	147	57	Alive	ASIAN	30	TUMOR FREE	1	1	0	No	85.78	POST	4	0	0	0	9	0	0	0	0	0	0	0	0	0	0	2

327	68	155	52	Alive	ASIAN	0	TUMOR FREE	1	1	0	No	82.33	POST	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	2
329	62	152	65	Alive	ASIAN	33	TUMOR FREE	1	1	0	No	72.67	POST	5	0	0	0	22	0	0	0	1	0	0	1	0	1	1	1
330	55	168	112	Alive	OR OTHER	65	TUMOR FREE	1	1	0	No	73.19	POST	8	0	0	0	8	0	0	1	0	0	0	0	0	0	0	2
151	47	163	53	Alive	WHITE	50	TUMOR FREE	1	1	0	No	40.08	POST	3	0	0	0	26	0	0	0	0	0	0	0	0	0	0	2
283	58	165	109	Alive	WHITE	52	TUMOR FREE	1	1	0	No	52.6	POST	5	0	0	0	16	1	0	1	0	0	0	0	0	0	0	2
284	89	130	56	Alive	WHITE	45	TUMOR FREE	1	1	0	No	52.79	POST	5	0	0	0	11	0	0	1	1	1	0	1	0	1	1	1
285	57	165	121	Alive	WHITE	0	TUMOR FREE	1	1	-1	No	56.54	POST	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	2
321	76	165	59	Alive	WHITE	0	TUMOR FREE	1	1	0	No	96.42	POST	1	0	0	0	10	0	0	0	0	0	0	0	0	0	0	2
328	54	170	114	Alive	OR OTHER	0	TUMOR FREE	1	1	0	No	81.5	PERI	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	2
336	55	160	93	Alive	ASIAN	0	TUMOR FREE	1	1	0	No	66.43	POST	11	0	0	0	14	0	0	0	0	0	0	0	0	0	0	2
147	52	168	132	Dead	WHITE	50	WITH TUMOR	1	1	0	No	61.04	POST	4	0	0	0	10	0	0	0	1	0	0	1	0	1	1	1
217	53	165	52	Alive	WHITE	69	TUMOR FREE	1	1	-1	No	59.79	PERI	4	0	0	0	20	1	0	0	0	0	1	1	0	4	2	3
218	55	0	0	Alive	WHITE	7	TUMOR FREE	1	1	-1	No	185.64	POST	0	0	0	0	13	1	0	0	0	0	0	0	0	0	0	2
221	65	159	124	Alive	AFRICAN A	8	WITH TUMOR	1	1	-1	No	75.66	POST	7	0	0	0	6	1	0	0	2	0	2	1	1	2	2	3
232	64	157	71	Alive	AFRICAN A	29	TUMOR FREE	1	1	-1	No	19.55	POST	0	0	0	0	3	1	0	0	0	1	0	0	0	0	4	2
235	59	162	63	Alive	WHITE	54	TUMOR FREE	1	1	-1	No	18.4	POST	5	0	0	0	8	1	0	0	0	0	0	0	0	0	0	2
236	61	172	84	Alive	WHITE	0	TUMOR FREE	1	1	-1	No	17.35	POST	0	0	0	0	0	1	0	1	1	0	0	1	0	1	1	1
279	56	170	144	Alive	WHITE	20	TUMOR FREE	1	1	0	No	51.51	PRE	3	0	0	0	3	0	0	1	1	0	0	1	0	1	1	1
324	59	170	99	Alive	OR OTHER	5	TUMOR FREE	1	1	0	No	92.81	POST	3	0	0	0	34	0	0	0	1	1	1	1	-1	2	2	8
339	59	161	58	Alive	WHITE	30	WITH TUMOR	1	2	0	No	29.93	POST	19	0	0	0	51	0	0	0	0	0	0	0	1	0	0	7
341	57	165	117	Alive	WHITE	12	TUMOR FREE	1	1	0	No	0.56	IND	10	0	0	0	48	0	0	0	1	0	0	1	0	1	1	1
343	34	170	94	Alive	WHITE	5	TUMOR FREE	1	1	0	No	28.65	PRE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
344	58	170	77	Alive	WHITE	30	TUMOR FREE	1	1	0	No	23.92	POST	0	0	0	0	23	0	0	0	0	0	0	0	0	0	0	2
345	80	154	51	Alive	WHITE	11	TUMOR FREE	1	3	0	No	22.54	POST	12	0	0	0	26	0	0	0	0	-1	-1	0	-1	0	6	9
346	60	150	82	Alive	WHITE	10	TUMOR FREE	1	1	0	No	1.74	POST	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
347	49	178	122	Alive	WHITE	13	TUMOR FREE	1	1	0	Yes	4.43	PERI	0	0	0	0	45	0	0	0	1	0	0	1	0	1	1	1
348	87	150	60	Alive	WHITE	0	TUMOR FREE	1	1	0	No	21.98	POST	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
20	76	150	57	Alive	WHITE	50	TUMOR FREE	1	1	0	No	73.82	POST	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	2
90	69	160	99	Alive	WHITE	12	TUMOR FREE	1	1	0	No	48	POST	6	0	0	0	13	0	0	0	0	0	0	0	0	0	0	2
102	82	161	72	Alive	WHITE	17	WITH TUMOR	1	1	0	Yes	93.92	POST	10	0	0	0	32	14	0	14	2	0	0	-2	0	1	6	2
108	33	166	100	Alive	AFRICAN A	71	TUMOR FREE	1	1	-1	No	27.1	PRE	13	1	0	1	40	0	0	0	0	0	0	0	0	0	0	2
144	60	160	98	Alive	DIAN OR A	50	WITH TUMOR	1	1	0	No	27.37	POST	4	0	0	0	20	0	0	0	0	0	0	0	0	0	0	2
146	63	161	112	Alive	WHITE	50	WITH TUMOR	1	1	-1	No	60.61	POST	4	0	0	0	21	0	0	0	0	0	0	0	0	0	0	2
201	63	168	119	Alive	WHITE	12	TUMOR FREE	1	1	-1	No	55.26	POST	0	0	0	0	14	1	0	0	1	0	0	1	0	1	1	1

219	71	160	77	Alive	AFRICAN A	0	TUMOR FREE	1	1	-1	No	113.27	POST	0	0	0	0	10	1	0	0	0	0	0	0	0	0	2
220	67	0	101	Alive	WHITE	0	TUMOR FREE	1	1	-1	No	88.27	POST	0	0	0	0	15	1	0	0	0	0	0	0	0	0	2
222	64	0	134	Alive	WHITE	9	TUMOR FREE	1	1	-1	No	54.4	POST	0	0	0	0	17	1	0	0	0	0	0	0	0	0	2
225	69	174	90	Alive	WHITE	5	TUMOR FREE	1	1	-1	No	61.63	POST	0	0	0	0	46	1	0	0	0	0	0	0	-1	0	0
226	62	177	114	Alive	WHITE	13	TUMOR FREE	1	1	-1	No	85.71	POST	0	0	0	0	18	1	0	0	0	0	0	0	0	0	2
228	64	161	88	Alive	WHITE	34	TUMOR FREE	1	1	-1	No	32.06	POST	0	0	0	0	8	1	0	1	1	0	0	1	0	1	1
229	51	168	89	Alive	AFRICAN A	60	TUMOR FREE	1	1	-1	No	77.17	POST	12	0	0	0	6	1	0	0	1	1	1	1	0	2	2
230	63	157	118	Alive	WHITE	17	TUMOR FREE	1	1	-1	No	73.23	POST	2	0	0	0	28	1	0	0	1	0	0	1	0	1	1
231	74	152	61	Alive	AFRICAN A	23	TUMOR FREE	1	1	-1	No	16.72	POST	5	0	0	0	16	1	0	1	0	0	0	0	0	0	2
233	61	163	102	Alive	WHITE	52	TUMOR FREE	1	1	-1	No	19.48	POST	0	0	0	0	17	1	0	0	0	2	0	0	0	4	2
237	57	165	80	Alive	WHITE	48	TUMOR FREE	1	1	-1	No	16.56	POST	0	0	0	0	0	1	0	0	0	0	0	0	0	0	2
280	53	162	147	Alive	WHITE	50	TUMOR FREE	1	1	0	No	51.02	PRE	2	0	0	0	6	0	0	1	2	0	0	2	0	1	1
281	77	157	93	Alive	WHITE	50	TUMOR FREE	1	1	0	No	51.97	POST	2	0	0	0	10	0	0	1	0	0	0	0	0	0	2
302	61	163	0	Alive	WHITE	39.4	TUMOR FREE	1	3	0	No	34.95	POST	32	29	0	29	14	9	0	9	0	0	0	0	0	0	2
305	74	154	59	Dead	WHITE	39.4	TUMOR FREE	1	3	0	No	30.98	POST	9	0	0	0	24	0	0	0	1	0	1	1	-2	2	2
307	47	166	93	Alive	AFRICAN A	17	TUMOR FREE	1	1	0	No	24.97	PERI	5	0	0	0	25	0	0	0	2	0	0	2	0	1	1
332	55	157	83	Alive	ASIAN	5.6	TUMOR FREE	1	1	0	No	89.52	POST	8	0	0	0	22	0	0	0	0	0	0	0	0	0	2
333	48	170	102	Alive	OR OTHER	36.7	TUMOR FREE	1	1	0	No	63.53	PRE	1	0	0	0	26	0	0	0	0	1	0	0	0	4	2
334	68	157	85	Alive	OR OTHER	75	TUMOR FREE	1	1	0	No	83.57	POST	5	0	0	0	13	0	0	0	0	-1	1	0	-1	0	0
337	60	168	82	Alive	WHITE	0	TUMOR FREE	1	1	0	No	16.82	PERI	0	0	0	0	0	0	0	0	1	0	0	1	0	1	1
338	75	165	110	Alive	WHITE	30	TUMOR FREE	1	1	0	Yes	19.78	POST	0	0	0	0	18	0	0	0	0	0	0	0	0	0	2
340	90	155	69	Alive	WHITE	60	TUMOR FREE	1	1	0	No	8.54	POST	0	0	0	0	0	0	0	0	1	0	1	1	0	2	2	3
342	54	164	62	Alive	WHITE	40	TUMOR FREE	1	1	0	No	4.86	POST	19	0	0	0	27	0	0	0	0	0	0	0	0	0	2
349	68	160	69	Alive	WHITE	20	TUMOR FREE	1	2	0	No	39.45	POST	16	0	0	0	38	0	0	0	0	-1	0	0	0	0	2
350	58	165	126	Alive	WHITE	38	TUMOR FREE	1	1	0	No	30.06	POST	8	0	0	0	34	1	0	1	0	0	0	0	0	0	2
351	45	169	72	Alive	WHITE	0	TUMOR FREE	1	2	0	No	29.86	PRE	16	0	0	0	29	0	0	0	0	0	0	0	0	0	2
352	72	160	125	Dead	WHITE	30	TUMOR FREE	1	1	0	No	1.91	POST	5	0	0	0	28	0	0	0	0	0	0	0	0	0	2
353	70	172	122	Alive	WHITE	20	TUMOR FREE	1	1	0	No	3.68	POST	13	0	0	0	29	0	0	0	1	0	0	1	0	1	1
354	78	159	71	Alive	WHITE	70	TUMOR FREE	1	1	0	No	29.3	POST	9	0	0	0	21	0	0	0	0	0	0	0	0	0	2
288	61	162	75	Alive	AFRICAN A	28	TUMOR FREE	1	1	0	No	13.76	NA	6	0	0	0	23	0	0	1	1	0	0	1	0	1	1
289	65	160	61	Alive	WHITE	80	NA	1	1	0	No	14.72	NA	3	0	0	0	4	0	0	1	0	0	0	0	0	0	2
290	74	162	84	Alive	WHITE	50	TUMOR FREE	1	1	0	No	13.99	POST	2	0	0	0	11	0	0	1	0	0	0	0	0	0	2
291	53	163	116	Alive	WHITE	39.4	TUMOR FREE	1	1	0	No	15.44	PERI	0	0	0	0	5	0	0	1	0	0	0	0	0	0	2
292	72	163	96	Alive	WHITE	13	NA	1	1	0	No	15.44	NA	5	0	0	0	25	0	0	1	1	1	1	-1	-1	2	5	6

313	39	165	109	Dead	OR OTHER	67	WITH TUMOR	1	1	0	No	46.91	PRE	2	0	0	1	2	0	0	2	0	0	0	0	0	0	0	2
356	50	169	109	Alive	WHITE	12	TUMOR FREE	1	1	0	No	32.56	PERI	38	0	0	0	53	2	0	2	0	0	0	0	0	0	0	2
357	64	181	87	Alive	WHITE	25	TUMOR FREE	1	1	0	No	0.66	POST	13	0	0	0	30	0	0	0	1	0	0	1	0	1	1	1
358	70	152	72	Alive	WHITE	60	TUMOR FREE	1	1	0	No	15.7	POST	20	0	0	0	52	0	0	0	0	0	0	0	0	0	0	2
359	67	162	76	Alive	WHITE	35	TUMOR FREE	1	1	0	No	36.07	POST	19	0	0	0	49	0	0	0	0	1	1	0	0	4	4	5
360	63	170	102	Alive	WHITE	45	WITH TUMOR	1	1	0	No	37.29	POST	23	0	0	0	29	0	0	0	0	0	0	0	-1	0	0	0
361	64	159	86	Alive	WHITE	56	TUMOR FREE	1	1	0	No	36.33	POST	12	0	0	0	18	0	0	0	0	0	0	0	0	0	0	2
362	49	164	151	Alive	WHITE	86	TUMOR FREE	1	1	0	No	37.48	PERI	0	0	0	0	34	0	0	0	0	0	0	0	0	0	0	2
363	73	154	60	Alive	WHITE	25	TUMOR FREE	1	1	0	No	36.76	POST	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
365	74	173	75	Dead	WHITE	100	WITH TUMOR	1	2	0	No	27.4	POST	15	0	0	0	30	0	0	0	0	0	0	0	2	0	0	7
366	62	163	84	Dead	WHITE	66	TUMOR FREE	1	2	0	No	11.5	POST	22	0	0	0	29	0	0	0	1	0	-1	-1	-1	1	6	9
367	61	174	130	Alive	WHITE	100	TUMOR FREE	1	1	0	No	29.89	POST	13	0	0	0	50	0	0	0	0	0	0	0	0	0	0	2
368	51	155	89	Alive	WHITE	42	TUMOR FREE	1	1	0	No	31.04	PERI	10	0	0	0	49	0	0	0	0	0	0	0	-1	0	0	0
369	44	169	102	Alive	WHITE	12	TUMOR FREE	1	1	0	No	33.61	POST	0	0	0	0	0	0	0	0	0	1	0	0	0	0	4	2
370	54	172	117	Alive	WHITE	77	TUMOR FREE	1	1	0	No	10.64	POST	19	0	0	0	74	0	0	0	0	0	0	0	0	0	0	2
371	39	159	105	Alive	WHITE	19	TUMOR FREE	1	1	0	No	0.69	IND	0	0	0	0	49	0	0	0	0	0	0	0	0	0	0	2
372	70	162	59	Alive	WHITE	100	TUMOR FREE	1	2	0	No	18.66	POST	20	0	0	0	35	1	0	1	1	-1	2	1	1	2	2	3
374	54	160	101	Alive	WHITE	18	TUMOR FREE	1	1	0	No	3.65	POST	26	0	0	0	46	0	0	0	0	0	0	0	0	0	0	2
375	56	156	87	Alive	WHITE	8	TUMOR FREE	1	1	0	No	17.48	POST	6	0	0	0	20	0	0	0	0	0	0	0	0	0	0	2
376	51	157	65	Alive	WHITE	8	TUMOR FREE	1	1	0	No	19.51	IND	9	0	0	0	19	0	0	0	0	0	0	0	0	0	0	2
378	67	167	94	Alive	WHITE	13	TUMOR FREE	1	1	0	No	30.12	POST	33	0	0	0	57	0	0	0	1	0	0	1	0	1	1	1
379	70	152	53	Alive	WHITE	93	TUMOR FREE	1	1	0	No	18	POST	15	0	0	0	45	0	0	0	1	0	0	1	0	1	1	1
381	59	176	102	Alive	WHITE	41	TUMOR FREE	1	1	0	No	1.58	POST	23	0	0	0	40	0	0	0	1	0	0	1	0	1	1	1
382	69	162	110	Alive	WHITE	8	TUMOR FREE	1	1	0	No	7	POST	5	0	0	0	42	0	0	0	1	0	0	1	0	1	1	1
383	78	161	81	Alive	WHITE	14	TUMOR FREE	1	1	0	No	0.99	POST	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	2
384	58	155	68	Alive	WHITE	67	TUMOR FREE	1	1	0	No	30.42	POST	36	0	0	0	23	0	0	0	0	0	0	0	0	0	0	2
385	68	158	79	Alive	WHITE	21	TUMOR FREE	1	1	0	No	1.28	POST	0	0	0	0	42	0	0	0	0	0	0	0	0	0	0	2
386	61	167	96	Alive	WHITE	57	TUMOR FREE	1	1	0	No	27.99	POST	0	0	0	0	28	0	0	0	0	0	0	0	0	0	0	2
387	74	161	97	Alive	WHITE	0	TUMOR FREE	1	1	0	No	28.06	POST	0	0	0	0	0	1	0	1	0	2	1	0	0	4	4	5
388	81	172	123	Alive	WHITE	22	TUMOR FREE	1	1	0	No	26.35	POST	0	0	0	0	34	0	0	0	1	0	0	1	0	1	1	1
389	56	163	142	Alive	WHITE	24	TUMOR FREE	1	1	0	No	10.22	POST	0	0	0	0	14	0	0	0	1	0	0	1	0	1	1	1
390	46	166	134	Alive	NA	0	TUMOR FREE	1	1	0	No	1.51	PERI	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	2
391	54	169	68	Alive	WHITE	12	TUMOR FREE	1	1	0	No	1.77	PERI	0	0	0	0	29	0	0	0	0	0	0	0	0	0	0	2
392	58	165	126	Dead	WHITE	5	TUMOR FREE	1	1	0	No	11.14	PRE	0	0	0	0	11	0	0	0	1	0	0	1	0	1	1	1

393	59	160	109	Alive	WHITE	18	TUMOR FREE	1	1	0	No	24.28	POST	0	0	0	0	52	0	0	0	0	0	0	0	0	0	0	2
394	67	159	83	Alive	WHITE	25	TUMOR FREE	1	1	0	No	2.33	POST	0	0	0	0	24	0	0	0	0	0	0	0	0	0	0	2
395	53	168	87	Alive	WHITE	100	TUMOR FREE	1	1	0	No	22.9	POST	9	4	0	4	3	0	0	0	0	2	0	0	0	0	4	2
421	64	173	107	Alive	WHITE	15.6	TUMOR FREE	1	1	0	No	16.29	NA	2	0	0	0	18	0	0	0	1	0	0	1	0	1	1	1
548	63	152	72	Dead	WHITE	67	NA	2	4	1	No	0	POST	5	3	0	3	6	1	0	1	2	0	2	1	-1	2	2	8
549	69	159	59	Dead	WHITE	82	WITH TUMOR	2	4	0	No	0	POST	0	0	0	0	1	0	0	0	1	1	1	1	-1	2	2	8
550	68	146	49	Alive	WHITE	28	TUMOR FREE	2	4	-1	No	0	POST	9	0	0	0	32	0	0	0	1	0	1	1	0	2	2	3
551	61	158	84	Dead	WHITE	6	WITH TUMOR	2	4	0	No	0	POST	3	3	0	3	0	0	0	0	0	0	1	1	1	4	2	3
552	62	152	70	Dead	WHITE	40	WITH TUMOR	2	4	0	No	0	POST	3	0	0	0	18	0	0	0	1	1	1	1	-1	2	2	8
553	71	152	76	Alive	WHITE	28	WITH TUMOR	2	4	0	Yes	0	POST	0	0	0	0	8	1	0	1	1	0	1	1	0	2	2	3
554	67	166	78	Dead	AFRICAN A	94	WITH TUMOR	2	4	-1	No	0	POST	10	5	0	5	13	9	0	9	0	1	0	0	-1	0	4	0
555	65	155	60	Dead	WHITE	60	WITH TUMOR	2	4	0	No	0	POST	5	0	0	0	12	0	0	0	0	1	0	1	0	0	3	1
556	69	172	56	Alive	WHITE	69	TUMOR FREE	2	4	-1	Yes	0	POST	0	0	0	0	13	3	0	3	0	1	1	1	2	4	2	5
557	54	160	58	Dead	WHITE	7	WITH TUMOR	2	4	0	No	0	POST	7	4	0	4	1	0	0	0	1	0	1	1	1	2	2	3
558	62	152	55	Alive	WHITE	0	TUMOR FREE	2	4	-1	No	0	POST	6	0	0	0	10	0	0	0	1	-1	1	1	1	2	2	3
559	62	155	88	Dead	WHITE	71	WITH TUMOR	2	4	0	No	0	POST	0	0	0	0	2	0	0	0	1	2	1	1	-1	2	2	8
560	65	172	73	Dead	AFRICAN A	67	WITH TUMOR	2	4	-1	No	0	POST	2	0	0	0	3	0	0	0	1	-1	-1	1	-1	1	1	9
561	59	163	86	Alive	WHITE	18	TUMOR FREE	2	4	0	No	0	NA	7	0	0	0	19	0	12	1	1	2	0	1	-1	1	1	8
562	83	164	82	Dead	WHITE	39.4	WITH TUMOR	2	4	0	No	0	POST	1	0	0	0	15	0	14	1	1	1	1	1	2	2	2	3
563	81	0	0	Dead	WHITE	39.4	WITH TUMOR	2	4	0	Yes	0	POST	5	0	0	0	2	0	0	1	0	0	0	2	0	0	0	1
564	84	0	0	Dead	WHITE	39.4	WITH TUMOR	2	4	0	Yes	0	POST	3	1	0	0	13	5	0	1	-1	1	-2	0	1	3	3	7
565	51	157	167	Dead	WHITE	39.4	TUMOR FREE	2	4	0	No	0	PERI	5	0	0	0	4	0	0	1	1	0	1	1	1	2	2	3
566	77	162	76	Alive	WHITE	40	WITH TUMOR	2	4	0	No	0	POST	5	0	0	0	7	3	0	1	0	1	-1	1	-1	3	3	4
567	65	157	70	Dead	WHITE	72	WITH TUMOR	2	4	-1	No	0	POST	7	0	0	0	6	2	2	2	0	0	1	0	0	4	0	2
568	87	140	42	Dead	WHITE	100	WITH TUMOR	2	4	-1	No	0	POST	0	0	0	0	0	1	0	1	-1	1	1	-1	1	4	4	5
569	72	157	64	Dead	WHITE	83	WITH TUMOR	2	4	0	No	0	POST	0	0	0	0	2	2	2	2	1	1	0	1	-1	1	1	8
570	69	152	95	Alive	WHITE	0	TUMOR FREE	2	4	0	No	0	POST	5	0	0	0	10	0	0	0	1	0	1	1	-1	2	2	8
571	71	160	99	Dead	WHITE	43	WITH TUMOR	2	4	0	No	0	POST	8	0	0	0	21	0	0	0	0	-1	0	0	0	0	0	2
572	82	156	60	Dead	WHITE	0	WITH TUMOR	2	4	0	No	0	POST	7	0	0	0	17	0	0	0	0	1	-1	1	1	3	3	7
573	73	155	92	Alive	WHITE	63	TUMOR FREE	2	4	0	Yes	0	POST	11	0	0	0	14	0	0	0	0	1	0	0	-1	0	4	0
574	68	158	63	Alive	WHITE	33	TUMOR FREE	2	4	0	No	0	POST	0	0	0	0	20	0	3	3	1	-1	1	1	1	2	2	3
575	66	138	119	Alive	AFRICAN A	33	WITH TUMOR	2	4	0	No	0	POST	0	0	0	0	0	1	0	1	1	-1	1	1	-1	2	2	8
576	88	161	57	Dead	WHITE	50	WITH TUMOR	2	4	0	Yes	0	POST	0	0	0	0	1	1	0	1	-1	1	-1	0	1	3	3	7
577	76	145	75	Alive	WHITE	8	TUMOR FREE	2	4	0	No	0	POST	6	6	0	6	12	7	0	7	-1	-1	-1	-1	1	0	6	7

579	77	162	70	Dead	WHITE	22.2	WITH TUMOR	2	4	-1	No	0	POST	3	1	0	1	2	1	0	1	2	0	2	1	-1	2	2	8
580	70	166	77	Dead	AFRICAN A	59	WITH TUMOR	2	4	0	No	0	POST	4	0	0	0	5	0	0	0	1	1	1	1	-1	2	2	8
581	65	162	90	Alive	WHITE	0	WITH TUMOR	2	4	-1	No	0	POST	5	0	0	0	20	0	0	0	1	1	0	1	0	1	1	1
582	61	148	68	Alive	AFRICAN A	8.3	NA	2	4	0	No	0	POST	9	0	0	0	7	1	0	0	1	1	0	1	1	1	1	1
583	69	161	95	Alive	AFRICAN A	33	TUMOR FREE	2	4	-1	No	0	POST	10	0	0	0	7	0	0	0	1	1	1	1	1	2	2	3
584	68	157	63	Dead	WHITE	4.2	WITH TUMOR	2	4	-1	No	0	POST	9	0	0	0	17	0	0	0	1	2	1	1	-1	2	2	8
585	72	170	79	Alive	AFRICAN A	39.4	TUMOR FREE	2	4	0	No	0	POST	0	0	0	0	16	0	0	0	1	-1	0	1	0	1	1	1
586	61	155	66	Alive	WHITE	39.4	TUMOR FREE	2	4	0	No	0	POST	2	0	0	0	19	0	0	0	0	-1	1	1	1	0	2	3
592	67	170	67	Dead	AFRICAN A	100	WITH TUMOR	2	4	0	No	0	POST	1	1	0	1	0	0	0	0	0	1	0	0	1	0	4	7
593	90	150	44	Dead	ASIAN	51	TUMOR FREE	2	4	0	No	0	POST	0	0	0	0	0	0	0	0	0	0	1	0	1	4	0	5
595	77	160	72	Dead	WHITE	100	WITH TUMOR	2	4	0	No	0	POST	0	0	0	0	0	0	0	0	2	2	1	0	-1	2	5	6
596	59	159	83	Alive	WHITE	39.4	TUMOR FREE	2	4	0	No	0	POST	21	0	0	0	24	0	0	0	0	1	0	0	1	0	4	7
597	61	157	83	Alive	NA	0	TUMOR FREE	2	4	-1	No	0	POST	11	0	0	0	24	0	0	0	0	1	0	-2	0	0	4	2
598	59	157	75	Alive	WHITE	39.4	TUMOR FREE	2	4	0	No	0	POST	16	0	0	0	34	0	0	0	1	0	1	1	0	2	2	3
599	90	162	77	Dead	WHITE	100	WITH TUMOR	2	4	-1	No	0	POST	10	7	0	7	18	7	0	7	1	0	0	1	-1	1	1	1
600	63	154	83	Dead	WHITE	10.5	WITH TUMOR	2	4	0	No	0	POST	0	0	0	0	0	1	0	1	2	0	2	1	-1	2	2	8
601	74	170	81	Dead	WHITE	54.5	WITH TUMOR	2	4	0	No	0	POST	4	0	0	0	22	0	0	0	1	0	0	1	0	1	1	1
602	67	160	73	Alive	WHITE	90	WITH TUMOR	2	4	0	Yes	0	POST	2	0	0	0	0	1	0	1	1	0	0	1	-1	1	1	1
603	55	157	50	Alive	ASIAN	33	TUMOR FREE	2	4	0	No	0	POST	0	0	0	0	6	0	0	1	1	0	-1	-1	0	1	6	9