Applying Machine Learning to Breast Cancer Gene Expression Data to Predict Survival Likelihood

Pegah Tavangar

Thesis submitted to the University of Ottawa

in partial Fulfillment of the requirements for the

Master of Science in Chemistry and Biomolecular Sciences

Department of Chemistry and Biomolecular Sciences
Faculty of Science
University of Ottawa

Abstract

Analyzing the expression level of thousands of genes will provide additional information beneficial in improving cancer therapy or synthesizing a new drug. In this project, the expression of 48807 genes from primary human breast tumors cells was analyzed. Humans cannot make sense of such a large volume of gene expression data from each person. Therefore, we used Machine Learning as an automated system that can learn from the data and be able to predict results from the data. This project presents the use of Machine Learning to predict the likelihood of survival in breast cancer patients using gene expression profiling. Machine Learning techniques, such as Logistic Regression, Support Vector Machines, Random Forest, and different Feature Selection techniques were used to find essential genes that lead to breast cancer or help a patient to live longer. This project describes the evaluation of different Machine Learning algorithms to classify breast cancer tumors into two groups of high and low survival.

Acknowledgments

I would like to thank Dr. Jonathan Lee for providing me the opportunity to work with him on an exciting project. I would like to recognize the invaluable counsel that you all provided during my research. It was my honor to work with some other professors in the Faculty of Medicine, such as Dr. Mathieu Lavallée-Adam and Dr. Theodore Perkins. Thank you for the support, advice, and friendship over the years. Their valuable guidance and advice over the past years significantly shaped my work. Finally, I would like to thank all my friends and family who have supported me throughout my graduate studies.

Table of contents:

Contents

List of Abbreviations

SVM	Support vector machine
RFE	Recursive feature elimination
mRNA	Messenger RNA
ER	Estrogen receptor
PR	Progesterone receptor
HER2	human epidermal growth factor receptor 2
tRNA	Transfer RNA
snRNA	Small nuclear RNA
RNA	Ribonucleic acid
cDNA	Complementary DNA
PCR	Polymerase chain reaction
Lasso	Least Absolute Shrinkage and Selection Operator

List of Figures

List of Tables

1: Introduction

1.1: Cancer

The studies show that breast cancer is a second common cancer worldwide[1]. At the mild level of cancer, cancer cells are those cells that cannot follow the reasonable control that the body remains on all healthy cells in our body[1]. There are billions and billions of cells in our body, and they have different functions. If something goes wrong and particular cells lose control of the standard mechanisms, they will continue to grow, and they may spread in an unusual way, which is called cancer[1].

Many cancer cells together would lead to a tumor, cancer is a malignant tumor, and it called malignant since cancer cells can invade other organs or spread to other tissues[1]. Which can be very dangerous and life-threatening and occur anywhere in the body[1]. Studies have shown, cancer that occurs in one individual could be different from cancer in another person[1].

Cancer treatment is very complicated. One of the main reasons is that cancer cells are very different from each other and are not homogeneous[1]. There might be different slight variations in the cancer cells, and that is why it is hard to control or treat cancer[1]. Sometimes it can shrink by 70%. Since there is a different subtype of that cancer that is going to require a different kind of treatment[1]. There are three primary therapies for breast cancer such as surgery, hormonal therapy, and chemotherapy[1]. Chemotherapy is a common treatment method for breast cancer[1]. Doctors normally use Chemotherapy after surgery for patients with early-stage breast cancer to decrease the risk of having cancer back and tumors[1]. Besides, it is one of the essential treatment for metastatic breast cancer. Studies show single chemotherapy drugs are less effective than a combination of them for breast cancer treatment[1]. The other most common therapy is surgery. The types of surgery mainly depend on the size of the tumor, location of the tumor and also if the spread of cancer reached the lymph nodes[1]. Hormonal therapy is also used for breast cancer therapy to cure hormone receptor-positive breast cancer[1].

1.1.1: Breast Cancer

Breast cancer is one of the most common cancers among women; approximately one in eight women will develop breast cancer throughout their lifetime [2]. Most breast cancers are detected on mammograms even before the patient notices it [1]. However, there are some prevalent symptoms such as a lump dimpling in the nipple, not symmetric breast, redness of the skin, things that are warm to touch [3].

Most breast cancers are not hereditary, which means there is not a clear genetic link to most breast cancer; however, there is a small proportion of breast cancer, approximately 5 to 10% of women with breast cancer do have a genetic predisposition for breast cancer, which means that a gene can be inherited from their mother or father [4]. Therefore, there is a higher probability of developing breast cancer over their lifetime [5]. There are some other risks that we cannot change such as age and personal factors such as, starting periods before age 12, menopause after age 55, being overweight, using hormone replacement therapy, and drinking alcohol [4].

Many studies show that studying different molecular subtypes of breast cancer may lead to better planning in the treatment and development of new therapies to cure breast cancer [6]. The complexity of different subtypes can be detected by the use of genetic information form the tumor cells[6]. There are 5 major molecular subtypes such as Luminal A, Luminal B, Triple-negative/basal-like, HER2-enriched and fifth Normal-like. Luminal A breast cancer is hormone-receptor-positive, which could be estrogen-receptor and/or progesterone-receptor positive and HER2 negative[6]. Based on the previous studies Luminal A breast cancer subtypes have low levels of the protein Ki-67, which regulate the growth of cancer cells. They also have the best prognosis and grow slowly[6]. Luminal B breast cancer is hormone-receptor-positive, which means it is estrogen-receptor and/or progesterone-receptor positive, and either HER2 positive or HER2 negative[7]. Besides, levels of Ki-67 are high in this subtype and normally grow faster than luminal A cancers[7].

The third breast cancer type is called Triple-negative/basal-like which is estrogen-receptor and progesterone-receptor negative and HER2 negative[7]. Studies Show women with BRCA1 gene mutations commonly have this type of cancer. Opposite to Triple-negative/basal-like, HER2-enriched breast cancer type is estrogen-receptor and progesterone-receptor negative and HER2 positive[7]. HER2-enriched may have a worse prognosis and generally grow faster than luminal cancers. However, they can be successfully cured with targeted therapies aimed at the HER2 protein[6]. The fifth type is called Normal-like breast cancer that is similar to luminal A disease. It is estrogen-receptor and/or progesterone-receptor positive and HER2 negative, and also has low levels of the protein Ki-67[6]. However, compared to luminal A cancer's prognosis, normal-like breast cancer has a slightly worse prognosis[7].

Estrogen receptor (ER), progesterone receptor (PR), and HER2 are important in breast cancer development. ER and PR are both hormones receptors that can found on the surface of cells such as breast cells, and they are responsible for taking in hormones signals that induce cell growth [8]. Human epidermal growth factor receptor 2 (HER2) controls breast cell division, growth, and repair of the cells. Consequently, it has a significant role in the expansion of breast cancer [3]. If the HER2 gene generates abnormal copies of itself, that can lead to producing too many HER2 receptors and, as a result, activate uncontrolled cell growth [8]. Investigation revealed that HER2-positive is more prone to metastasize and is more likely to reappear, which means HER2-positive breast cancer lower rate of survival than HER2 negative breast cancer [9].

1.1.2: Illumina bead chip:

Who we are, how we built all of these pieces of information embedded in our genetic code[10]. There are genetic variations across multiple individuals which analyzing and understanding them can help scientists to gain helpful insight to improve human health and disease [11].

Nowadays, gene expression analysis was done with microarrays[12]. All types of cells have the same type and number of genes, but the expression varies from different types of cells or different times during different developmental stages[13]. Microarray or DNA chip consists of a microscope slide with thousands of tiny grooves in defined positions and all the genes of a particular organism are placed in that different spot[13].

Gene expression means analyzing the different expression changes in RNA level expression.on another world; gene expression is the process that the information of the genes is used to produce gene products[13]. These products are often proteins, but in non-protein coding genes such as transfer RNA (tRNA) or small nuclear RNA (snRNA) genes, the product is a functional RNA, and those mRNA will produce proteins[13]. Therefore, by measuring the mRNA content, or the amount of mRNA in a given time, the gene expression can be measured. The amount of mRNA present in a given time inside the cell is called the transcriptome[13].

There are two stages for analyzing gene expression. The first stage is to produce the DNA chip, and the second stage is to measure the transcriptome[13].to produce the DNA a chip; all the genes in our desires organism should be amplified using PCR ((Polymerase Chain Reaction)[13]. Then, the DNA should be denatured to get single-stranded DNA and place it in a DNA chip. After this step, the chip is available for any gene expression analysis[13].

In the second stage, the total amount of mRNA of the cell at a particular time is extracted and then reverse transcript process applied by using the enzyme called reverse transcriptase to produce cDNA. cDNA is single-stranded DNA, and it will be acting as a control[13]. After that, the same type of cells is selected to extract the same RNA and finally similarly get the cDNA. The second step is to tag the DNA to be able to find it in the chip[13]. So, the total cDNA extracted from the cell will be tagged with the dye that is called cy3. Non-control cDNA should also be tagged with a different dye, and then microarray can be used for the gene expression analysis[13]. Next, tagged control cDNA transfers into a microarray, and in the end, there are two single-stranded DNA that will be available at each spot to hybridize[13]. Then DNA chip will be placed into an analyzer. The analyzer scans the genes to find the colors to detect the type of genes that are more or less expressing[13].

Illumina Infinium assay is a DNA analysis tool to measure these genetic variations within each individual, and it also helps to explore how it causes different traits and diseases[11]. To be able to do this Illumina Infinium assay with bead chip technology is used [11]. In Illumina Infinium assay, the genotype of locus uses two different color readouts, one color for each allele[10]. The surface of each bead chip, millions of genome types, can be assayed at once. These silica beads are placed in microwell which is coated with multiple copies of an oligonucleotide probe that targets a specific locus in the genome [11]. Each probe bind to its complementary sequence in the sample DNA as each DNA fragments pass over the chip, stopping one base before the locus of interest[11]. Allele specificity is conferred by a single base extension which incorporates one of the four labeled nucleotides[11].

Once the laser gets excited, the nucleotide label emits a signal which is detected by Illumina scanner. Intensity values for each color give information about the ratio of a given locus[10]. The data then analyzed by using a luminous genome Studio software to evaluate copy number variation across the genome[10]. When the essay data from several individuals are plotted distinct patterns emerge this approach is advancing the understanding of genetics [11].

Although the Microarray or DNA chip technique can simultaneously measure the expression levels of a large number of genes, there are some limitation applies to this method[13]. As an example, the sensitivity of the microarray analysis is insufficient for the detection of low-abundance genes[14]. Besides, the false-positive rate in this method is high since cross-hybridization can generate non-specific labeling. Furthermore, the dynamic range of detection to measure expression levels of genes is limited[14].

As a result, microarray data normally requires more accurate and sensitive quantitative analysis[14]. As a result, the next-generation sequencing approach developed for more accurate and sensitive analysis that is called RNA sequencing. RNA sequencing can provide a rapid and deep investigation of the transcriptome, for different species[14]. This approach has some important advantages to microarray analysis, for example, RNA-Seq technology does not require species or transcript specific probes, and also it can detect gene fusions, single nucleotide variants, small insertions, or deletions, in general, different changes that microarrays can not detect[14]. Besides, RNA-Seq technology has higher specificity and sensitivity with compare to a microarray, so it can easily detect genes with low expression or high expression[14]. It can also provide detection of rare and low-abundance transcripts[14].

1.1.3: Machine Learning

We are surrounded by many data that are generated by humans and also by computers such as phones, pictures, videos, etc. The Machine Learning goal is to provide meaningful patterns from all of the given data. Typically, humans manually analyze the data to find a pattern or alter systems to the changes[12]. However, since the volume of data can be very large, humans cannot make sense of it and that is when Machine learning can be used as an automated system to identify patterns in data[15]. One example of usage of Machine Learning is Google search. Each time we use Google search, we are using a system that has many Machine Learning systems at its core to understand the text of the query to modify the results based on personal interests, for example knowing what results to show the first[15].

Machine Learning application is broad that include image recognition, text recognition, fraud detection, and speech detection.it can also be applied to science fields such as skin cancer detection[16]. Therefore, Machine Learning can be used to make human tasks better, more comfortable, or to do tasks that cannot be achieved manually[16].

1.1.4: Supervised and Unsupervised Learning

Supervising a model means teaching a model with some knowledge so that it can predict future instances[17]. To teach a model, the model should be trained based on a labeled dataset to be able to predict the output of sample data. There are two types of supervised learning, which are called Classification and regression; however, in unsupervised learning, the model works on its own to predict the new information that may not be clear or visible to the human eye[17]. Therefore, unsupervised learning uses different Machine Learning algorithms that predict unlabeled data[17]. Generally, with unsupervised learning (figure 1) is mostly used in clustering or grouping of unlabeled sample data. On the other hand, in supervised learning, the model knows what kind of data is existed[17].

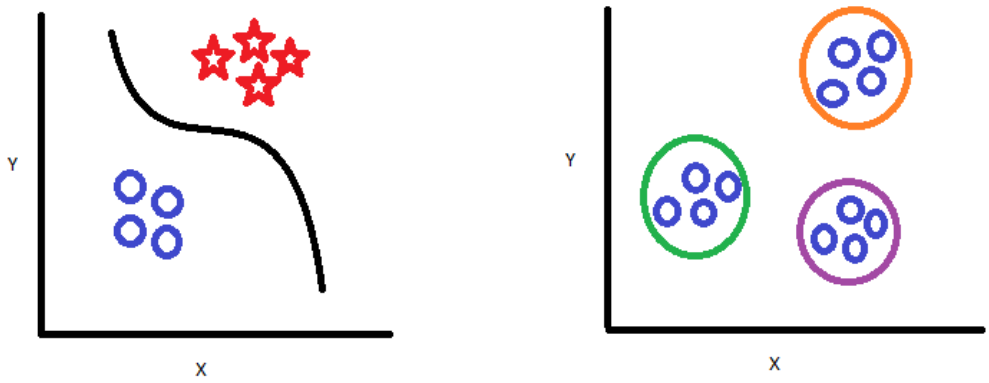


Figure 1. Unsupervised learning model the right-pane works on its own to discover information and it is mainly required unlabeled data. On the other hand, a supervised learning model the left–pane learns from labeled training data to predict outcomes for unforeseen data. Adapted from reference [18].

1:2: Machine learning algorithms

1.2.1: Logistic Regression

Logistic Regression is one of the foremost utilized statistical procedures applied by the researcher for the investigation of binary or multiple variable reaction information[19]19. So in Logistic Regression, the hypothesis is going to look as below[19].

Θ = Parameter

x = Input data

b = intercept

Linear Regression hypothesis:

$$h_{\theta}(x) = \theta^{T}x + b$$

Logistic Regression hypothesis:

$$h_{\theta}(x) = g(\theta^{T}x + b) = 1/1 + e^{-\theta Tx}$$

1.2.2: Loss Function

So to learn parameters for any model, a training set of m training examples are given, and the goal is to find parameters θ and b so that the predictions you have based on the training set will be close to the true labels Y (target output) that you got in the training[19]. Therefore, the Loss Function or error function is used to measure how well the algorithm is doing[19]. The Loss Function is always for a single training example and measures how well the algorithm works on the single training example; however, there is another term which is called Cost Function, which measures how well the entire training set is doing[19].

1.2.3: Random Forest

Random Forest is a classification algorithm that comprises of many decision trees. The decision tree has a flowchart structure which consists of a different number of internal node[20]. A decision tree asks a question and then classifies the data based on the answer [21]. In order to make a Random Forest "bootstrapped" dataset needs to be created [22]. The bootstrapped dataset has the same size as the original, but samples randomly selected from the original dataset into a bootstrapped dataset. Also, choosing the same sample more than once is possible [21]. The next step in Random Forest is the creation of multiple decision trees using the bootstrapped dataset. However, a random subset of descriptors or columns at each step is selected to make decision trees [22]. There are parameters in Random Forest that are called max features, and a different number of features can consider using the best set of split at each step of making a decision tree [21].

The classification can be categorized or numeric [23]. The first node in the decision tree is called the root node or root, and after the root node, there are internal nodes that are called nodes, which indicates a test on an attribute[21]. Therefore, the internal node has arrows pointing to them and also arrows pointing away from them[21]. The last nodes of decision trees are called leaf nodes that have arrows pointing to them. However, no arrows are pointing away from them[21].

In a simple decision tree, all the features are considering to split the root node, but in the Random Forest, the only subset of features is selected by using a max feature parameter [22]. After, choosing the best features for a root node and continue randomly selecting of remained features until reach to the leaf nodes [21], Figure 2 shows a simple Random Forest classifier with 4 different trees.

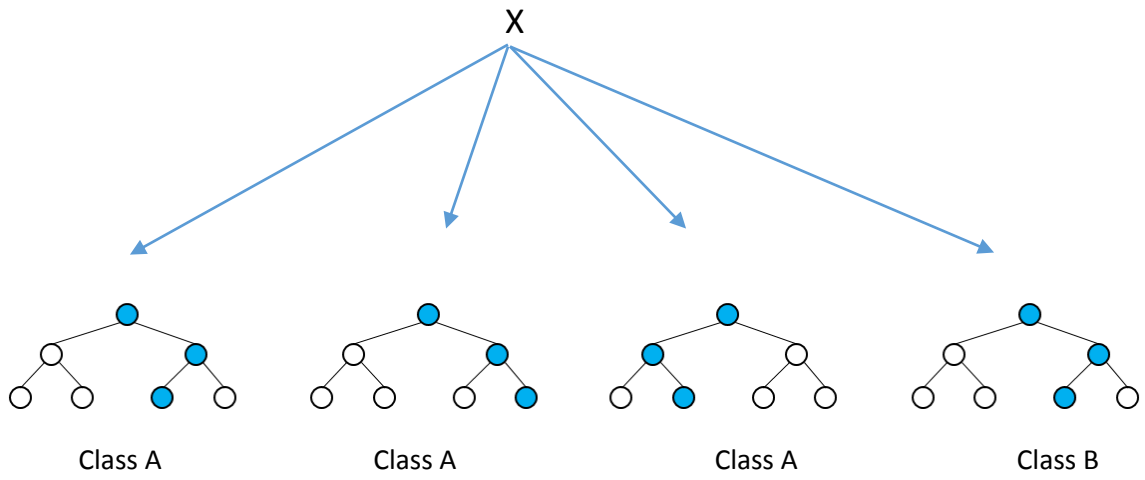


Figure 2. The Random Forest consists of a large number of decision trees that work as an ensemble. Each tree in the Random Forest involve in prediction and the class with the most votes determines the final prediction. Adapted from reference [24].

1.2.4: Support Vector Machines

Support vector machines are another type of supervised learning models that can be used for regression and classification problems [25]. The goal of using a Support Vector Machine is to find a decision boundary in an N-dimensional space in a way that classifies the samples in two distinct groups [26].

There are many possible decision boundaries to choose from, but the goal is to find a hyperplane that provides the maximum margin or distance between the data points of the two different classes [27]. Since the bigger, the margin gives more confidence for the classification of data points into two distinct groups [27]. The dimension of hyperplane depends on the number of features available so, if there is 2 number of features in total then the hyperplane will have a two-dimensional plane [26].

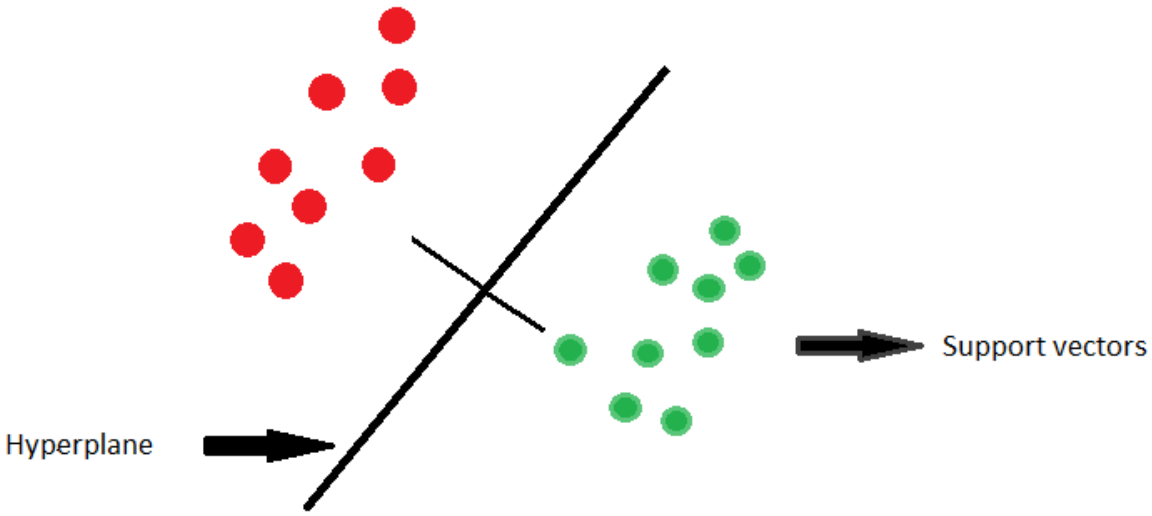


Figure 3. Hyperplane representation of the SVM algorithm. Adapted from reference [28].

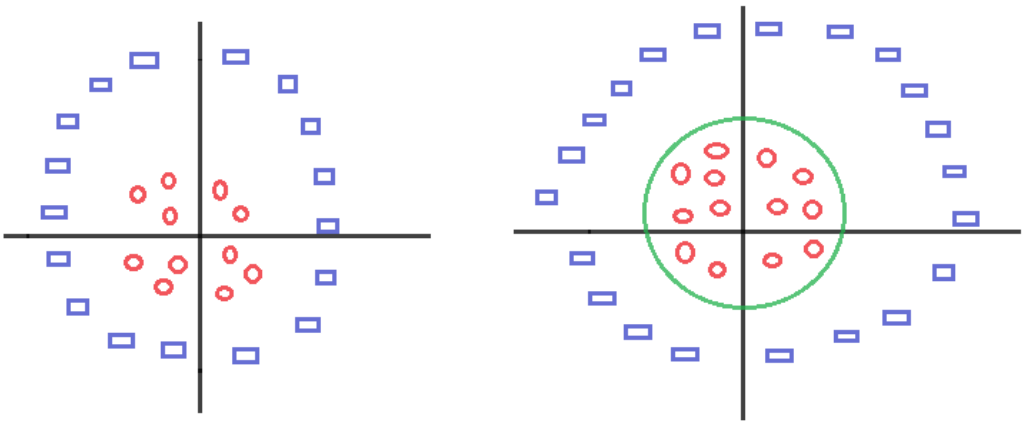


Figure 4. When the data is not linearly separable, SVM can be used to apply transformations to the data and map the data from the original space into a higher dimensional feature space. Adapted from reference [29].

SVM is a very beneficial algorithm especially when the separation of data points into separate classes cannot be possible with the existence dimension such as the example shown in figure 6 [30]. To separate figure 6 data sets into two classes; the SVM model is used to add one more dimension as the z-axis and transform it into a 3-dimensional plane as is shown in figure 7. This transformation is called the Kernel trick [30]. The kernel trick is a method to use linear classifiers for non-linear problems. Therefore, it transforms non-linearly separable data into linearly separable data [30].

Linear classifier depends on dot product between vectors [27].

$$K(X_i, X_j) = X_i * X_j = X_i^T * X_j$$

After mapping each data point into high dimensional space via some transformation Φ, then the dot products would become look like [27].

T = Transpose of a matrix

$$K(X_i, X_j) = \Phi(X_i)^T * \Phi(X_j)$$

K is the Kernel function, which is a similarity function that corresponds to an inner or dot product in some expanded feature space [30].

There are different types of the kernel, such as linear, polynomial, Gaussian, Radial basis function, and sigmoid [30]. Kernel function returns the inner product of two points into a sufficient feature space [27]. Some standard kernels used with SVM are shown below.

1.2.5: Linear Kernel

The Linear kernel is the simplest kernel function, and it is calculated by the inner product plus constant C, which is an optional parameter [31].

C = Constant

$$K(x, y) = x^T y + c$$

1.2.6: Polynomial kernel

The Polynomial kernel is very common in image processing problems and is calculated by the below equation[32]:

$$K(X_i, X_j) = (X_i * X_j + 1)^d$$

d = degree of the polynomial

1.2.7: Gaussian kernel

Where there is no advanced knowledge about the Gaussian data kernel is the best kernel to use, and it is measured by using Euclidean Distance of two vectors[32]. Parameter σ in the below equation shows the smoothness of the function [31].

The equation is shown below:

σ = Smoothness of the function

$$K(x, y) = exp(-||x - y||^2 / 2\sigma^2)$$

1.2.8: Radial Basis Function (RBF)

Radial Basis Function measures the similarity between vectors by calculation of squared norm of their distance, and this function has a bell-shaped curve[32].

Equation:

$$K(X_i, X_j) = exp(-\gamma ||X_i - X_j||^2)$$

γ = measure the similarity of vectors by the width of the bell-shaped curve[32].

1.2.9: Sigmoid kernel

The Sigmoid kernel is similar to the Sigmoid function in Logistic Regression, and it can be calculated by the equation below[32].

α and c are kernel parameters

$$K(x, y) = \tanh(\alpha x^T y + c)$$

1.2.10: Clustering

Clustering is a type of unsupervised learning, and it consists of grouping data [15]. Although the algorithm would not know much about the given data, it can split it into different groups [15]. There are two essential methods in Hierarchical Clustering[33]. The first one is called k means clustering, and the second was called a Hierarchical Cluster [15]. In K-means clustering, K means it is a number of clusters to identify [15]. In K-means clustering, k centroids are used to define clusters; therefore, to select the number of the cluster, the sum of distances of observations from their cluster centroids is calculated to choose the number of the cluster for given data set [15].

There are two different methods in Hierarchical Clustering, Agglomerative hierarchical clustering, and Divisive Hierarchical Clustering[34]. In Agglomerative clustering, each data point is initially in a cluster of its own, and at each step, the two closest clusters are combined into a single cluster[34]. Oppositely, in Divisive methods, all the data points are in a single cluster, to begin with, and recursively split the cluster[34].

There are two key steps to build a Hierarchical Clustering algorithm. The first step is calculating the similarity between two points, and one method to do that is called Euclidean distance[34]. Euclidean distance can be calculated by taking an average of two points, and the average is also a point in the Euclidean space[34]. The points that have the least distance are referred to as similar points, and therefore it merged. This step repeats this process until only a single cluster is left[34]. The sequence of merges is shown by a tree-like, a diagram that is called a dendrogram [34].

1.3: Evaluation Metrics

Evaluating the performance of the model with a different combination of parameters is an essential part, and there are different methods to use such as Accuracy, Confusion Metrix, Mean Squared Error, Mean Absolute Error, Root Mean Square Error, R-squared, F_1 score and Area Under the Curve [20].

1.3.1: Accuracy

Accuracy is defined as the number of correct predictions compared to the total number of input samples. In defining accuracy several parameters are used such as Confusion Metrix that indicates essential terms such as True Positives (when the both predicted, and the actual output is yes), True Negatives (when we predict no, and the actual output is no), False Positives (when we predict yes, and the actual output is no) and False Negatives (when we predict No and the actual output is yes)[20].

1.3.2: Area Under The Curve

This method is used for binary classification problems. AUC can be represented by plotting False Positive Rate as the x-axis and True Positive Rate as the y-axis [35].

True Positive Rate is also called Sensitivity and calculate as below [20].

True Positive Rate = True Positive /False Negative + True Positive

False Positive Rate is also called Specificity and calculate as below [20].

False Positive Rate = False Positive / False Positive + True Negative

In classification algorithms such as Logistic Regression, a threshold is chosen to binarize the data into two groups of zero or one [35]. By changing the threshold will lead to different Confusion Matrices. So instead of being confused by different Confusion Matrices to select the best threshold with the lower False Positive Rate, plotting a True positive Rate versus False Positive Rate, which is called Receiver Operator Characteristic (ROC) graphs can help to provide a simple way to summarize all of the vital information [20]. The AUC makes it easy to compare one ROC curve to another [35]. Therefore, Roc curve makes it easier to identify the best threshold for making a decision and AUC helps in selecting the best categorization method such as Logistic Regression versus Random Forest [35].

1.4: Cross-Validation

In Cross-Validation, data is split into a training and a test set. Cross-Validation has a different technique, but the famous one is called K-fold CV[36]. In K-fold CV data asset can further split into k number of the subset that is called fold[37]. It is arbitrary to choose K value. After that, the model fitted K times, and at each time, data trained on K-1 of the fold and the Kth fold is used to measure the performance of the model [37]. For example, if K is equal to 6, in the first iteration, the model train on the first five-folds and the evaluation occurs on the 6th fold, and in the second time the model trained on the first, second, third, fourth and sixth and evaluate on the fifth[37]. The procedure repeated 4 more times, and each time evaluation happened on a different fold. Lastly, the average performance on a different fold is calculated. An example of K-fold Cross-Validation is shown in below [38].

Cross-Validation is one of the most critical techniques to validate the stability of the machine learning model and how well the model can generalize to the new data[37]. It is very important to make sure that the model learns from informative patterns from the data correctly and does not pick the noise or irrelevant pattern[37]. Therefore, Cross-Validation will generate low bias and variance results to achieve better prediction performance[37].

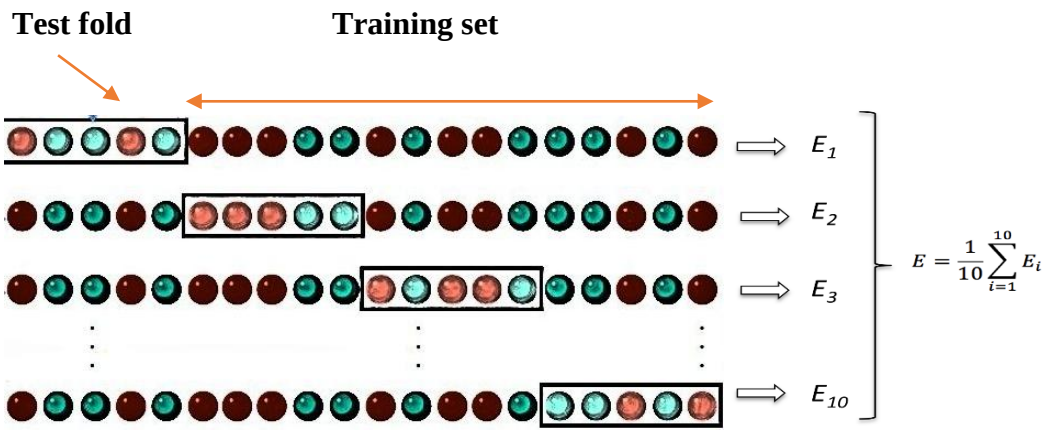


Figure 5. In K-Fold CV data set is split into a K number of sections or folds, where each fold is used as a testing set at some point. Adapted from reference[36].

1.5: Feature Selection

Feature Selection is a very crucial process to identify features (i.e. individual genes from an expression array) that play a crucial role in the prediction variable or output [39]. Having a large number of features can lead to overfitting and reduction in the accuracy of a model[40]. Choosing the right method of Feature Selection reduces the probability of making decisions based on noise, less misleading data, and high computational efficiency[40]. Feature Selection in Machine Learning hugely depends on the data and context of it[40]. There are many different methods of solution for Feature Selection. The best way to choose the best method is necessary to understand the different mechanism of each method and use when required[40].

Feature selection can provide insights about the features and, more importantly, their relative effects with the target variable[40]. Therefore, choosing the essential features can lead to having a higher accuracy prediction rate from the model and a lower error rate[40]. Also, having a lower number of features means there will be lower dimensionality in the system and, therefore, lower computational complexity[40]. In general, complex models with too many features are hard to decode, mainly when the features are very close or correlated with each other[40].

1.5.1: Recursive Feature Elimination

Recursive Feature Elimination (RFE) is one method in feature selection. RFE removes features recursively and only applies a model on remaining features that have a rank of 1 or most important ones so, it uses the model accuracy to measures which feature contributes more in predicting the output result [39].

1.5.2: Least Absolute Shrinkage and Selection Operator (Lasso)

Lasso is a fundamental unsupervised learning technique that is used to reduce the over-fitting problem and also help in feature selection as well. The Cost Function Lasso regression is shown as below [19].

m = Number of samples

$h_\theta(x^{(i)})$ = predicted output

$y^{(i)}$ = True output

$$J(\theta) = 1/2\,m\,[\sum (h_\theta(x^{(i)}) - y^{(i)})\,2 + \lambda \sum \theta^2_j]$$

The second term on the right of the above formula is a regularization term, and lambda is called the regularization parameter which controls, a trade-off between two different goals [19]. The first goal is for the first term of the formula, and the objective is to trade to fit the training data well, and the second goal is to keep the parameters small, which are captured by the second term by the regularization objective [19].

The second crucial step is to get an optimal number of clusters for hierarchical clustering. Since a dendrogram can visualize the steps of hierarchical clustering, a threshold can be set to only consider those clusters that have the lowest distance from each other[19]. This is how the number of clusters chooses by using a dendrogram in Hierarchical Clustering.

2: Materials and Methods

2.1: Dataset preparation

The data set comes from a genomic nature article [15]. In that study, an integrated analysis of gene expression for 48803 individual genes in a discovery and validation set of 997 and 995 primary breast tumors, were provided respectively[41]. In their work, over 2,000 clinically primary fresh-frozen breast cancer from tumor banks in the UK and Canada were collected [41]. After that, DNA and RNA different patients were isolated from samples and hybridized to the Affymetrix SNP 6.0 and Illumina HT-12 v3 platforms for genomic and transcriptional profiling, respectively[41]. Besides, different clinical features were collected in addition to the gene expression data, and those include; the size of the tumor, age, treatment, stage of the tumors, mutation types, etc. However, only 3 main clinical features such as progesterone receptor, estrogen receptor, human epidermal growth factor receptor, and survival of patients in days were selected among them.

In this project, we first applied Logistic Regression, Random Forest, SVM with all of the 48803 gene expression that includes noncoding and coding genes plus clinical features. Because of the low performance of different models, the number of features reduced to 770, including all 3 clinical features after applying feature selection models except with Lasso penalty. On the other hand, after using Lasso regularization with Logistic Regression the number of features reached 11 including all 3 clinical features.

The original discovery data set was shown as estrogen and progesterone receptor features as a plus or negative sign. Because our goal is to use binary classification the positive and negative were transformed to one and zero respectively. Besides, the target column represents the survival rates in continuous value in days and therefore they also transformed into binary values. Those that lived more than 6 years transformed into one and those that died less than 4 years transformed to zero.

Furthermore, since the gene expression data were very close from one patient to another patient, all the patients that were lived within 5 years were removed from the original data set to put a one-year gap between two distinct groups of high and low survivals and make it less confusing for the model when it comes to predicting whether a test data point is considered to be in the high or low survival group. After applying all the pre-processing steps like only considering those patients who died less than 4 years and lived for more than 6 years, the data set decreased to 533 from 997 data points.

2.2: Validation Test

In order to see how truly the chosen algorithms predict the output result, an unknown data set should be selected and check the performance of the model with the data that never seen before. Therefore, 557 new data as a validation test set was selected, and it applied with Logistic Regression, SVM, and Random Forest. Also, validation data pre-processing followed the same transformation steps that have been done in the training data set.

2.3: Logistic Regression performance

The logistic Regression algorithm can output its prediction, which is the estimate of output Y to be the probability of the chance[42].

Θ = Parameter

x = Input data

b = intercept

$$h_{\theta}(x) = g(\theta^{T}x + b) = 1 / 1 + e^{-\theta Tx}$$

$$g(z) = 1 / (1 + e^{-Z})$$

g (z) is called logistic or sigmoid function. If z is a huge number, then e^{-Z} will be very close to zero, so, sigmoid of z will be close to one[42]. Conversely, if z is very small, or it is a huge negative number, then sigmoid of z will be close to zero[42].

Sigmoid function (figure 7) has an S-shape cure and it always outputs a probability value between $0 < h\theta(x) < 1$[42]. In math, the hypothesis output is the probability that Y equals to 1 given X parametrized by theta[42].

$$h_{\theta}(x) = p(y = 1 / x ; \theta)$$

We can also compute the probability that Y is equal to zero[42].

$$h_{\theta}(x) = p(y = 0 / x ; \theta)$$

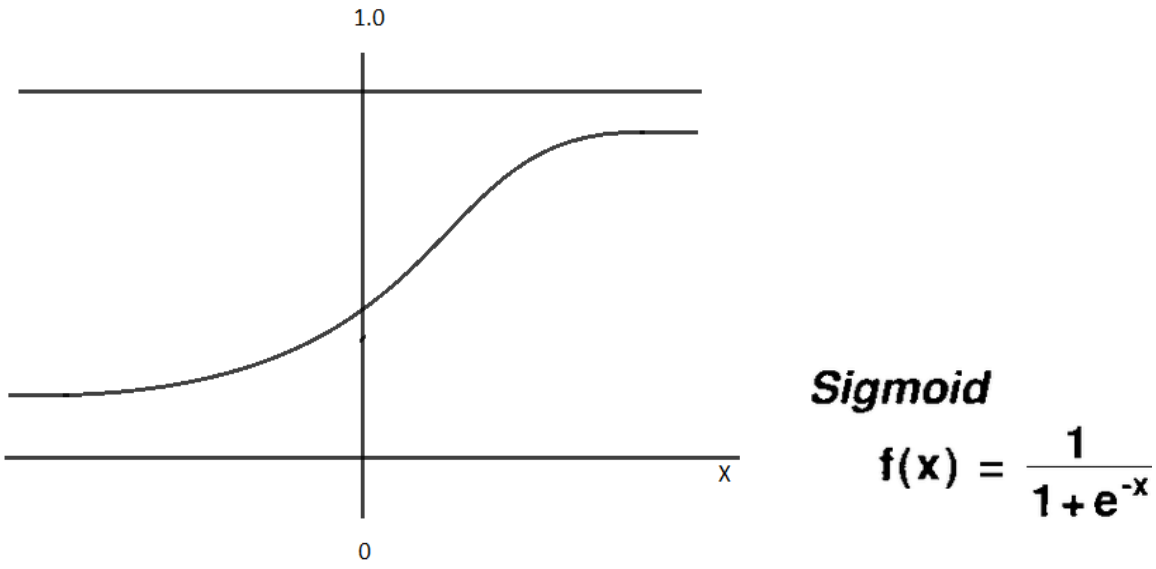


Figure 6. The hypothesis is going to predict output equal to one whenever X is greater or equal to 0 and predict zero when its less than zero. Adapted from reference[42].

2.3.1: Logistic Regression Cost Function

The Cost Function, which is applied the parameters θ and b, is going to be the average, one over m of the sum of the Loss Function applied to each of the training examples [43]. The Cost Function is used to measure the performance of a model for a given x value, and most of the linear regression problem its equal to the sum of the squared difference between predicted values and expected values as below [43].

m = Number of samples

$h_\theta(x^{(i)})$ = predicted output

$y^{(i)}$ = True output

$$J(\theta) = 1/2\,m \sum (h_\theta(x^{(i)}) - y^{(i)})^2$$

So Cost Function is the cost of parameters and measures how the model performs on the entire training set [43]. In Logistic Regression, the Cost Function formula is different from linear regression. In Logistic Regression, the Cost Function is shown below [43].

$$J(\theta) = -1/m \sum [y^{(i)} \log((h_\theta(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)}))]$$

That's because when we want to learn the best parameters, you will find optimization problem non-convex with the multiple global optimum instead of one global optimum. Therefore, an appropriate Cost Function makes it easier to optimize the parameters of a model [43].

Alpha is the learning rate and controls how big a step we take on each iteration of gradient descent[44]. So that hopefully, whether the initializing occurs left or on the right gradient descent will always move toward the global minimum[44]. In Logistic Regression, the Cost Function is a function of both θ and b, so the inner loop of gradient descent will be as follows[44].

2.4: Scikit-learn

Scikit-learn is a free Machine Learning library for the Python programming language, which consists of different functions and classes to do certain tasks [12]. Before applying k-fold Cross-Validation,sklearn.cross_validation.train_test_split package must be used to split the data set into the testing and training set [12].

Two main parameters need to be modified in sklearn.cross_validation.train_test_split package. The first one is called n-split, and the second one is random-state.n-split is used to set the proportion or rate of the training set to the test set, and therefore, the rate of 25% test set to 75% training set is chosen to apply different algorithms on the dataset [73]. The random-sate parameter is used to generate random-sampling. Fundamentally, if we do not specify the random_state, each time we executing the appropriate code, new train, and test datasets will be generated, which leads to different attributes at each node per running and different accuracy [73]. This is a way to Shuffle the features which can reduce the greediness of the algorithm. However, fixed integers value (0 or otherwise) is to make the outcome consistent across each time calls [73]. In this work, random-state specified to zero.

2.5: Random Forest as a Feature Selection method

2.5.1: Gini Impurity and Entropy

In the Random Forest, we randomly select features for the first node or the root node of the decision tree. For example, if there is a total of 4 features, Random Forest randomly selects 2 of them, and among that, only one of them selected[45]. This is when the Gini impurity or entropy method is used to select the best features in separating the samples [45]. Always a feature with lowest impurity or highest information gain will split the first [46]. Gini impurity is calculated how often a random sample from the data set will be incorrectly labeled if it is randomly labeled according to the distribution of labels in the subset[47]. Therefore, it shows the probability of classifying the data point incorrectly [46].

If there are N total classes and p (i) is equal to the probability of choosing a data point with class I, the Gini Impurity formula is equal to[46]:

$$Gini = \sum p(i) * (1 - p(i))$$

Entropy is another principle to split a node in a decision tree based on the given features. Entropy calculates the probability of disarrangements or disorder in a class of examples, and the equation of Entropy is shown below [46].

$$Entropy = -\sum p(i) \log p(i)$$

P (i) is a probability of a certain feature [46]

After calculation of the parent or output entropy from the above formula, it is then time to calculate the entropy of a child or different features and finally subtract the parent entropy from child entropy and choose that feature that has a less difference [47]. Hence, it means that the feature has the highest information gain or less impurity [46]. Making a new bootstrapped dataset and building a tree by only considering a subset of variables at each step is a repeat, and it arbitrary to choose a different number of trees by choosing different integer values for n_estimators parameters at Random Forest algorithm[48]. By repeating this step, we will end up with a large variety of trees, and that's what makes Random Forest as a very effective algorithm compare with a single decision tree [21].

In order to use the Random Forest, the data is run it down through the first tree and restores the prediction result and then moves to the next tree until it runs down all of the trees available in the Random Forest [48]. After that algorithm checks which option received more votes, for example, if we want to predict heart disease and there is two class patient with heart disease and without [47]. Then the number of votes that were collected based on each tree add together and only chose that class that has the highest votes [21].

2.5.2: Random Forest Hyperparameter tuning using Scikit learn library

In order to get a better result from a Random Forest algorithm, Hypperparamter tuning using Scikit learn library tool is very beneficial [45]. Hyperparameters are the parameters that need to be set by the data science before the process of model training [21]. In Random Forest, many hyperparameters include the number of features for each decision trees when splitting a node, numbers of trees, depth of the minimum tree number of a leaf node, etc [49].

Scikit learn is an open-source Machine Learning library that contains many tools and functionality for classification, regression, and other statistical models [47]. There are default hyperparameters for all statistical models whereas, it is not optimal for all the problems [49]. The selection of best hyperparameters depends on results on prediction, and the best technique is to try as many different combinations and evaluate the performance of the model at each time [47]. This method of selecting the best hyperparameters has one big problem which is fixable, and that is an over-fitting problem [49].

Each time the evaluation of the model occurs on only the training set and that can lead to a significant problem that it is called over-fitting [48]. Over-fitting or high variance means fitting the training set and have a hypothesis that passes through all of the data but fail to generalize to the new or unseen examples [47]. The Cross-Validation method is typically used to optimize hyperparameter [50].

It is very insufficient to try many hyperparameters by using K-fold CV. The best approach is using Scikit-learn's RandomizedSearchCV method [50]. By using the RandomizedSearchCV library grid of hyperparameter ranges can be defined but only randomly sample from the grid is chosen, and K-fold CV with each different combination applies on them [50]. The key hyperparameters in the Random Forest are listed below [50].

- n_estimators = number of trees considered in the Random Forest
- max_features = maximum of the number of features for splitting a node
- max_depth = maximum number of levels or length in each decision tree in Random Forest
- min_samples_split = minimum number of data points in each node before node get splitted
- min_samples_leaf = minimum number of data points placed in a leaf node
- Bootstrap = method for sampling the data points

After creating a parameter grid for each parameter mentioned above, the algorithm will select a different combination of them, and the benefit of random search is that only random of a combination of features will be selected [49]. In RandomizedSearchCV there is also a parameter is called n_iter which controls the number of different combinations to select [50]. Consequently, by choosing a large number, more iteration will occur, and a more comprehensive search of different values for each critical parameter will be covered [50].

2.6: Support Vector Machine Cost Function

The Cost Function of the Support Vector Machine is very similar to the Cost Function of the Logistic Regression[30]. SVM hypothesis predicts one When $\theta^T x \geq 0$ and predict 0 otherwise. The Cost Function is used in all the different algorithms to train and mainly optimize them. Minimization of the value of J (theta) will lead to the SVM accuracy in prediction[30]. In the below equation, the functions cost1 and cost0 corresponding to the cost where y=0 and the cost for where y=1[30]. The Cost for SVM is measured by the kernel (similarity) functions.

SVM Cost Function formula [25].

$h_\theta(x^{(i)})$ = predicted output

$y^{(i)}$ = True output

$$J(\theta) = \sum [y^{(i)} Cost_1 (h_\theta(x^{(i)})) + (1 - y^{(i)}) \cos t_0 (h_\theta(x^{(i)}))]$$

SVM has a regularization term so if regularization term is added the Cost Function would change to the below formula [25].

C = SVM parameter

$$J(\theta) = C [\sum [y^{(i)} Cost1 (h_\theta(x^{(i)})) + (1 - y^{(i)}) \cos t0 (h_\theta(x^{(i)}))] + ½ \sum \theta^2_j$$

C parameter in the above equation prioritizes the optimization term and regularization term. Choosing a significant value of C is the same as not using regularization term, and that means small margin hyperplane will be chosen by the optimization if it helps to achieve a better result in classifying the training points, but it is not going to generalize to the test set [25].

However, a small value of C leads to having a more substantial margin hyperplane even though that leads to misclassifying more points, which would lead to lower accuracy in the training set and higher accuracy result in the test set[27].

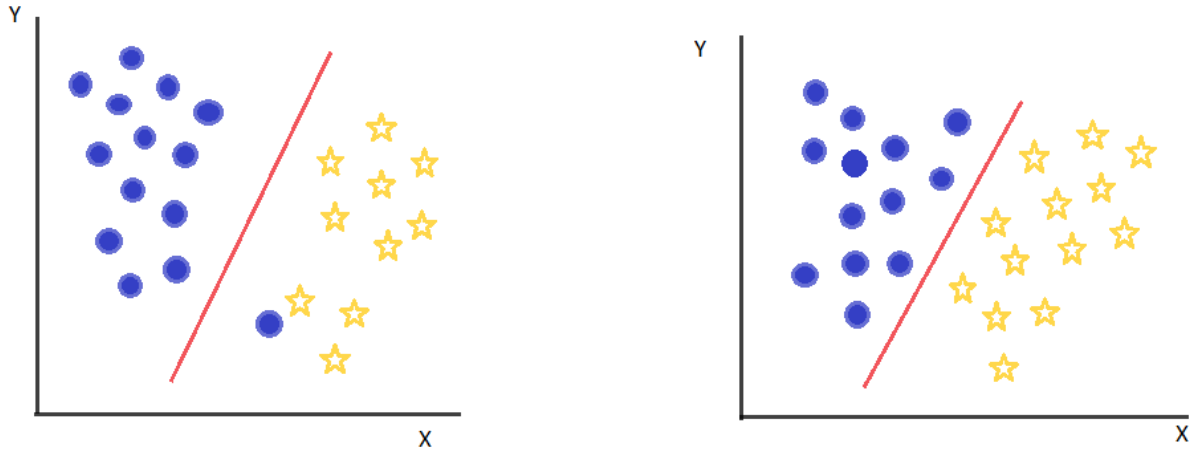


Figure 7. The left-pane show the low value of C and a high value of C on the right-pane. Adapted from reference [51].

There is another parameter in SVM which is called gamma, and it measures the influence of single training examples[27]. Gamma with low value means that every point has a far reach which means far away points considered where to draw a decision boundary and oppositely, a high value of gamma means each training example only has a close reach which means the decision boundary will depend mainly on the points very closest to it and that leads to ignoring the far point from the decision boundary[27].

2.6.1: Tuning Support Vector Machine parameters

In order to achieve better prediction, parameter tuning is required. Scikit learns packages provide GridSearchCV library for choosing the best C or gamma and also the best kernel (similarity function)[27]. Without using GridSearchCV loop over different values of parameters is required and run all the possible combination with Cross-Validation method which is not an efficient way to do[27]. By using GridSearchCV, a parameter grid can be set up, and then the algorithm will pass and return the best set of parameters and kernel to choose[27].

2.7: Feature selection

Feature Selection is the process of selecting those features that are effective and important to the prediction variable or output [52]. Feature Selection has a very crucial role in the simplicity, efficiency, and accuracy of the model [52].

2.7.1: Recursive Feature Elimination

Recursive elimination means to start from one feature then, add one by one attribute per loop, fit the model and calculate the accuracy of prediction to eliminate those features that have no contribution in increasing the model accuracy or reducing it [53].

In order to find the optimal number of features, Cross-Validation is used to combine with RFE. In this way, different feature subsets can get scored, and the best scoring features can be chosen [53].

2.7.2: Regularization parameter in Lasso

The regularization parameter lambda controls the tradeoff between these two goals, which will be fitting to training set well and the goal of keeping the parameters small[19]. Therefore, it keeps the hypothesis relatively simple to avoid overfitting. For select, of the best lambda value, a list with different range of numbers from 1×10^{-2} and 400 was chosen, and finally, the best value selected to calculate the final accuracy rate of the corresponding model.

2.8: Evaluation Metrics

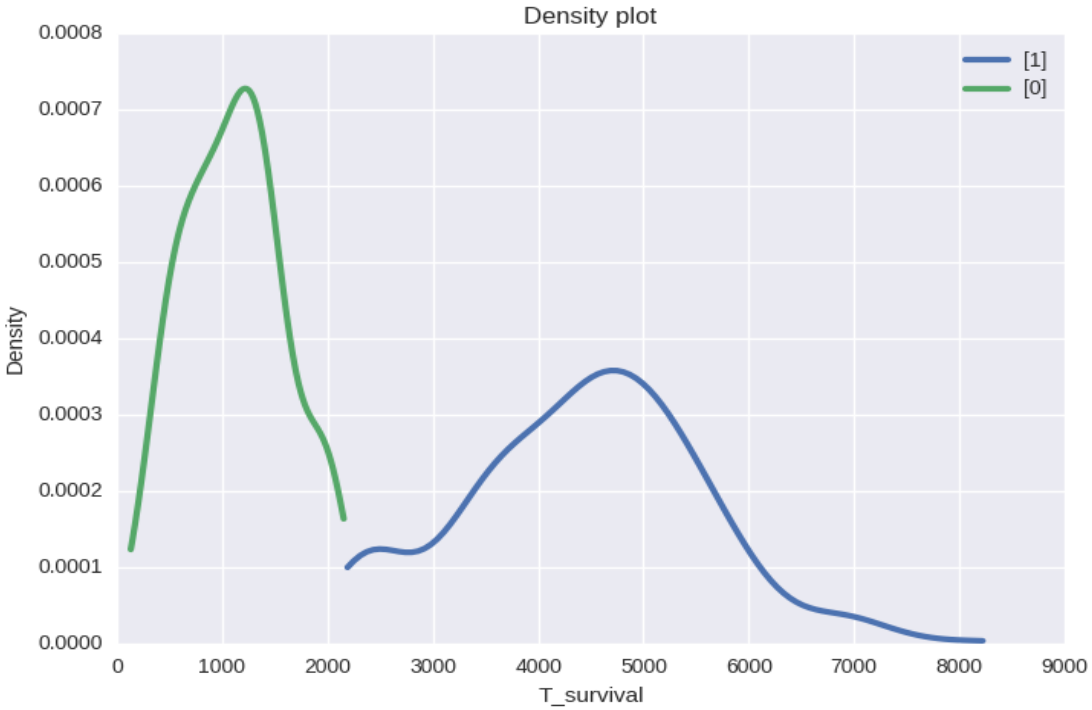
The type of evaluation metric used to evaluate the performance of different algorithms was accuracy, which means to calculate the number of correct predictions to the total number of input samples.

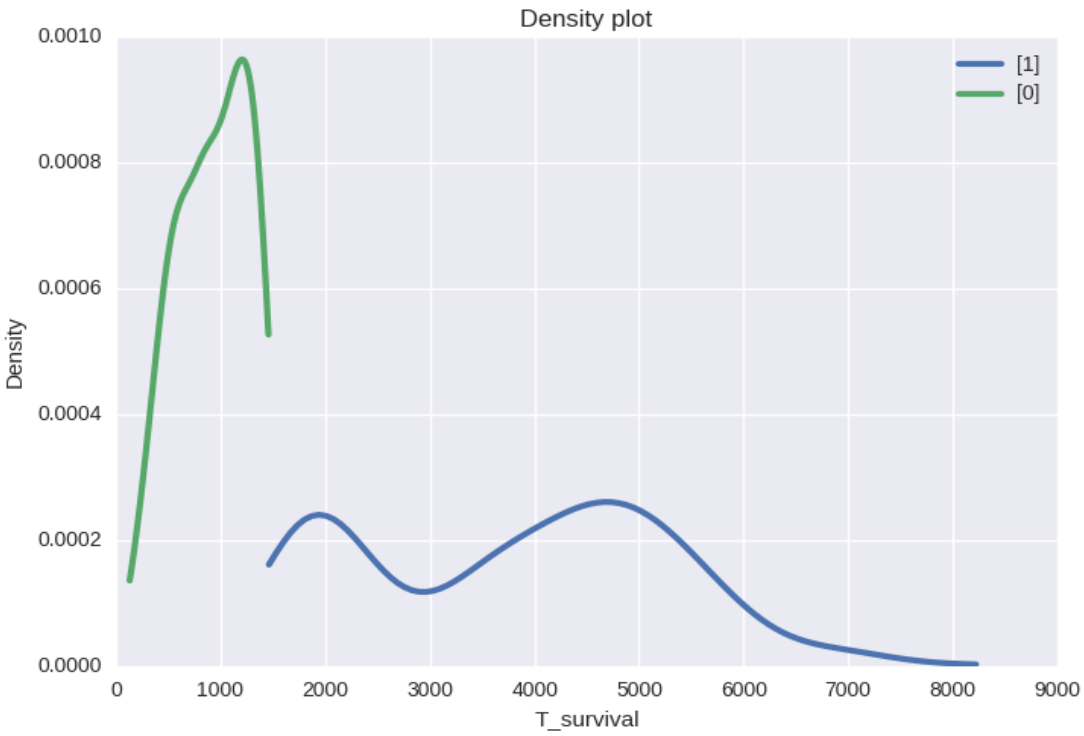
3. Results:

In this project, our goal is to combine gene expression data with some critical clinical data that would add more quality to the data set, such as estrogen receptors, progesterone receptor, human epidermal growth factor receptor 2 and then apply and evaluate the performance of different classification and feature selection algorithms to find optimal techniques for survival prediction in breast cancer.

The importance of feature collection to build a data set is the first and most important step in the Data Science field because the quality of predictions is limited by the information contained in the input data. So, if there are no relevant attributes or if the data are insufficient or unreliable, it can be difficult to make correct predictions even with an appropriate machine learning algorithm. Therefore, in this project or goal is to predict, if a patient with breast cancer will have a short or long term survival probability. Therefore, a binary classification problem should be used when the output labels in supervised learning problems are either zero or one [26]. The target is the longevity of patients that are continuous numerical data that needed to binarized into Zero and one.

Before binarizing the target feature, the probability density function was plotted to determine the best threshold to separated samples into two groups of zero and one. After trying multiple thresholds, it has been determined that patients who live more than 6 years and who died less than 6 years is the best way to classify samples because there would be no overlapping between long term and short term groups and also lead to almost 50/50 division of two groups which means the classifier will have an equal variety of the data set in two groups for train purposes. The below figures illustrate the probability density function plot with the different thresholds of 4, 6, and 7 years.


Density plot
[1]
[0]
Density
0.0008
0.0007
0.0006
0.0005
0.0004
0.0003
0.0002
0.0001
0.0000
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
T_survival


Density plot
[1]
[0]
Density
0.0010
0.0008
0.0006
0.0004
0.0002
0.0000
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
T_survival

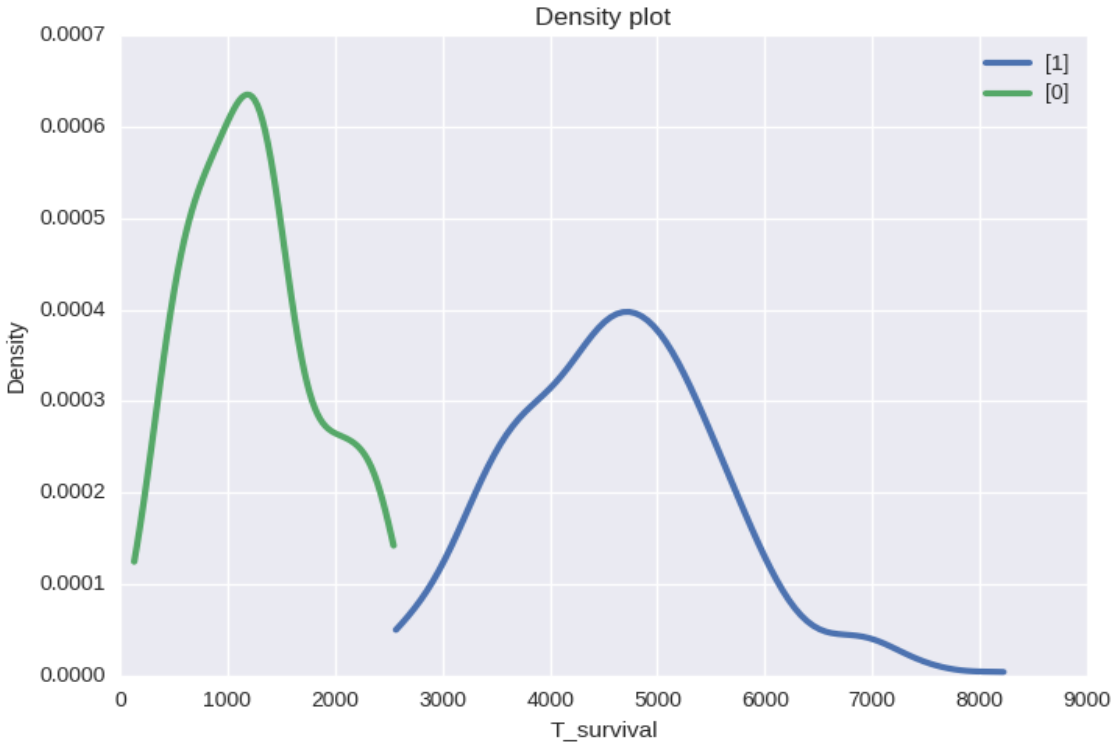


Figure 8. When the 6, 4 and 7-year threshold were selected respectively.

Furthermore, since the gene expression data were very close from one patient to another patient, all the patients that were lived within 5 years were removed from the original data set to put a one-year gap between two distinct groups of high and low survivals and make it less confusing for the model when it comes to predicting whether a test data point is considered to be in the high or low survival group. After applying all the pre-processing steps like only considering those patients who died less than 4 years and lived for more than 6 years, the data set decreased to 533 from 997 data points.

3.1: Logistic Regression

Logistic Regression was used to predict the probability of survival into two groups of long term and short term survival. The accuracy or number of correct predictions over the total number of predictions was equal to 73% after applying for the Logistic Regression model. 73% accuracy was achieved after applying one of the reliable ways to do a Cross-Validation technique that is called K-fold Cross-Validation, where the training set split into K subsets. Since 5 fold Cross-Validation was chosen, 5 times iteration over these 5 subsets applied, and the average of all different 5 accuracy was calculated.

3.2: Random Forest

The data set consists of a large number of gene expression features. This makes the model prone to overfitting problems, which leads to high accuracy in training examples and low accuracy with the test set. In order to overcome this problem, an ensemble algorithm called Random Forest regression was used. For that reason, the Scikit-learn RandomizedSearchCV library for the Random Forest model is used to choose the best hyperparameters for the Random Forest. In this method, the algorithm searches over different distributions of predetermined lists of values for Random Forest parameters and randomly selects a different combination[21]. Finally, that allows choosing the best hyperparameter value to be. As the name RandomizedSearchCV suggest this method is combined with Cross-Validation to achieve the best result[21].

After using RandomizedSearchCV the below is the result of best parameter selected for Random Forest model Bootstrap:False, min_samples_leaf: 2, n_estimator: 34, max_features: log2 (number of features) , min_samples_split: 3, max_depth: 39. The accuracy of the model reached the highest at 71 % with those parameters. In addition to the RandomizedSearchCV method, the SelectFromModel library from Scikit Learn was used to choose features that their importance is greater than the mean importance of all the features [54]. However, it can be altered by choosing a different threshold. Therefore, a list with different threshold was used to see which of them lead to the highest accuracy. The threshold that leads to the highest accuracy of 76 % is equal to 0.0019, with only129 features.

3.3: SVM

Support vector classification is another classification model that used to see if the data is separable in high dimensions rather than linearly [30]. The support vector model is combined with the GridSearchCV library to generates a grid of parameter values that can be specified by the parameter called param_grid [55]. Three main grids should be explored; the first one is the type of kernel to choose and then the C and gamma values [55]. The best parameters selected for C and gamma after trying a different range of value is equal to 0.001. Also, the linear kernel was chosen as the best kernel. After applying the 5-fold Cross-Validation, the total accuracy reached 72%.

3.4: Recursive Feature Elimination

Finally, after trying different algorithms and achieving a close range of accuracy in prediction, it was time to try a Feature Selection method to remove the unnecessary or weakest features and also overcome to overfitting problem to have a more accurate prediction. Therefore, RFE with Cross-Validation is chosen to apply with Logistic Regression and SVM model. The goal is to remove highly correlated features in the data set since they contain or provide the same information, and therefore RFE is a great way to rank each feature based on its importance and effect on model prediction [39]. After applying RFE, the number of features decreased from 48804 to 770. Based on the results, the accuracy is raised and reached 77 % using the Logistic Regression model and 76% with the SVM model.

Furthermore, the same 770 best features were used with the Random Forest model. After using Scikit-learn's RandomizedSearchCV library to choose the best parameters such as: bootstrap: False, min_samples_leaf: 2, n_estimators: 23, max_features: sqrt, min_samples_split: 5, max_depth: 31 the accuracy reached to 74%.In total, RFE increased the accuracy of prediction 4 percent in the Logistic Regression and SVM model. In the Random Forest, there was only a 3 percent increase inaccuracy.

3.5: Lasso

The problem with using RFE is that after the selection of essential features and applying Logistic Regression or SVM, both models are even more prone to overfitting problems than without using RFE. Despite reducing the number of features, the overfitting problem did not solve, because only the best features are selected, and that means the algorithm fits the training data set too well but failed to generalize well for the validation data set. Therefore, as shown in Table 1, the accuracy of the validation data set decreased 22 percent for logistics regression with only 23 selected features.

In order to overcome the overfitting problem and get higher accuracy in the validation Lasso regression analysis method was used instead of RFE. Lasso consists of a penalty factor that regulates the number of features that are maintained in a data set [19]. Besides, a combination of Lasso with a Cross-Validation method to select the penalty factor helps the model to generalize well to validation data samples [19]. Furthermore, Lasso's main difference to RFE is that Lasso provides not only the most essential features but output a useful set of selected predictors [19]. For that reason, I select a different range of lambda the penalty parameter and monitor how the accuracy would change, and in the end, the best accuracy with the least number of features was selected. The below table represents results after applying the Lasso penalty on the Logistic Regression model both on validation and discovery data.

As shown in Table 1, using Lasso leads to only a 1 percent error in the prediction of unseen or validation data, which is a significant improvement. All the results shown so far were when the threshold was set in such a way to separate patients into groups of high survivors who lived more than 6 years and low survivors who died less than 6 years.

In addition to that type of classifying samples into two groups of high and low survivals, the logistic classifier with lasso penalty was used again but with only considering patients who died less than 2 years and lived more than ten years. Finally, The accuracy of the Lasso penalty with Logistic Regression reached 88% accuracy with only 16 features for Discovery data and 86% with 19 features for validation data set. The reason for a significant jump inaccuracy is the selection of threshold of 2 and 10 years, and that's because of the 8 years gap between high and low survival instead of 1 year, and in this case, the gene expression between these two distinct groups is more distinguishable and detectable.

3.6: Feature Selection by using only Landmark 979 genes

Based on some studies, it is found out that there are about 979 crucial genes that can lead to identifying gene biomarkers for breast cancer survival [15]. Based on that study Logistic Regression with Lasso penalty with only 979 landmark genes that capture 82% of the information of the genome was tested, and it reached to highest accuracy of 63% with 4 features for discovery data [15]. Therefore, we conclude that those 979 landmark genes do not play a significant role in our data set.

Table 1. Different models and techniques used with both validation and discovery sets.

Name of the model and method	Validation accuracy %	Discovery accuracy%	Number of features in the validation set	Number of features in the discovery set
Logistic Regression with RFE	55	77	770	770
SVM with RFE	68	76	770	770
Random Forest	62	74	770	770
lasso with Logistic Regression	72	73	23	11
Random Forest with selecting threshold of 0.0019	-	76	-	129
Logistic Regression	-	73	-	48806
SVM	-	72	-	48806
Random Forest	-	71	-	48807

3.7: Heat map with Hierarchical Clustering

A heat map with Hierarchical Clustering in gene expression analysis is usually used to cluster rows and columns based on the similarity [33]. In the below figures, rows represent different genes or features that were selected after applying the Lasso penalty on the Logistic Regression model and the column represents different samples or tumor IDs. A heat map with Hierarchical Clustering helps to see any correlations in the dada set.columns can show what group of tumors express the same genes and row represents what groups of genes behave the same [33]. Figure 11 and figure 12 shows no separation between the different sets of tumors which may cause by overfitting problem. Logistic Regression model may be overfitted when 2 and 10 years threshold was selected and that forced the model to learn from all the noise and un relevant features to train the model, so once new data is tested it could not generalize well and therefore fail to do a correct prediction.

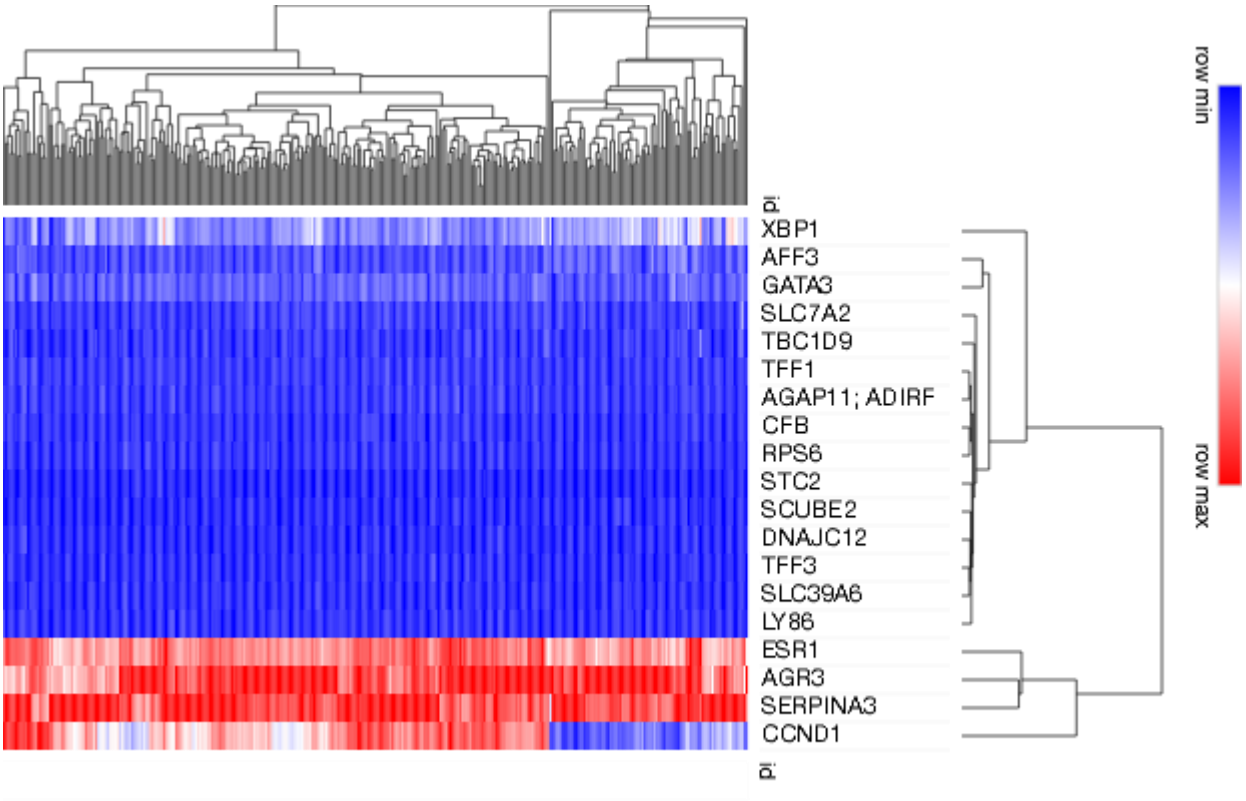


Figure 9. Heat map of Lasso penalty applied on the validation set with a threshold of 2 and 10 years. The above heat map shows, what type of genes act similarly to each other. The color and intensity of the boxes are used to shows changes in gene expression levels, for example, red represents up-regulated genes and green represents down-regulated genes.

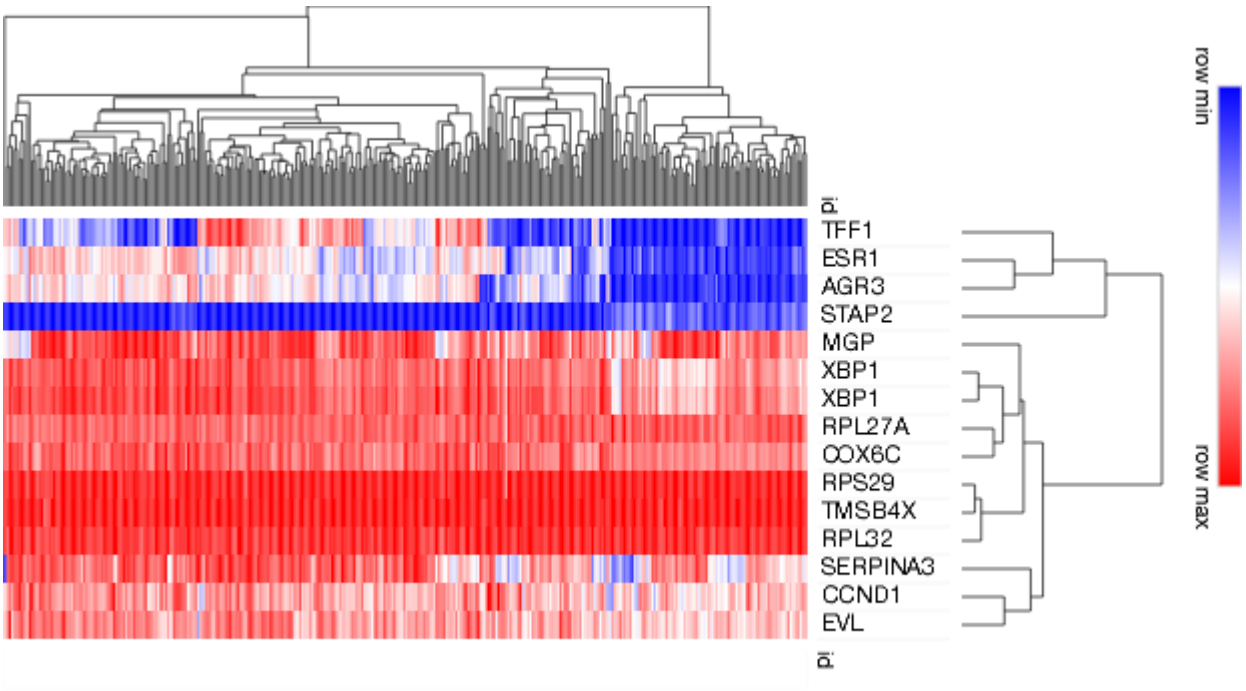


Figure 10. Heat map of Lasso penalty applied on the discovery set with a threshold of 2 and 10 years. The above heat map shows, what type of genes act similarly to each other. The color and intensity of the boxes are used to shows changes in gene expression levels, for example, red represents up-regulated genes and green represents down-regulated genes.

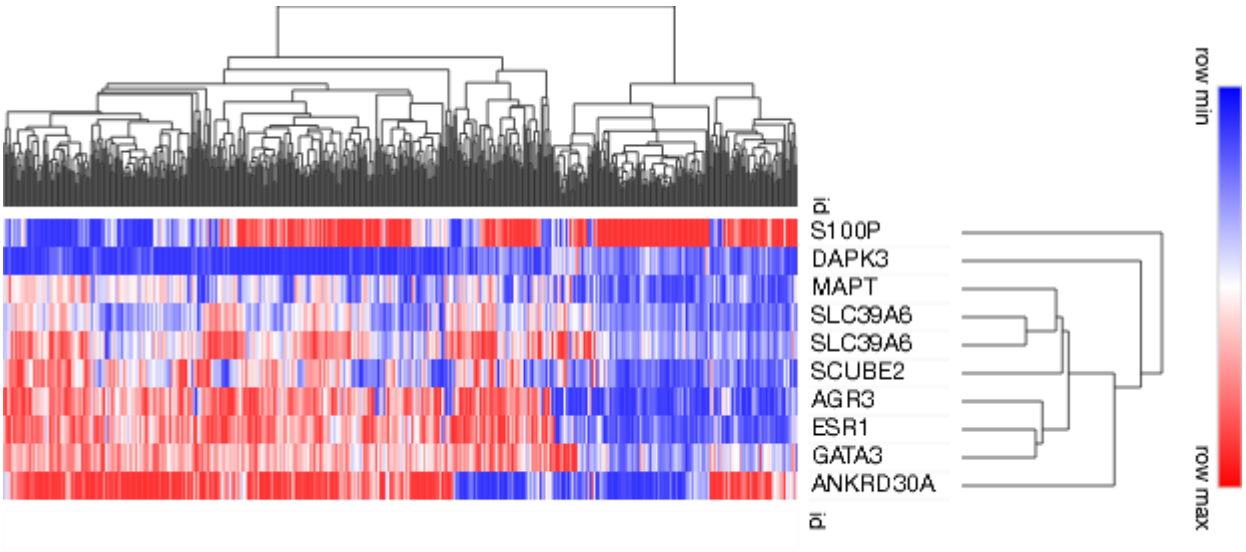


Figure 11. Heat map of Lasso penalty applied on the discovery set with a threshold of 4 and 6 years. The above heat map shows, what type of genes act similarly to each other. The color and intensity of the boxes are used to shows changes in gene expression levels, for example, red represents up-regulated genes and green represents down-regulated genes.

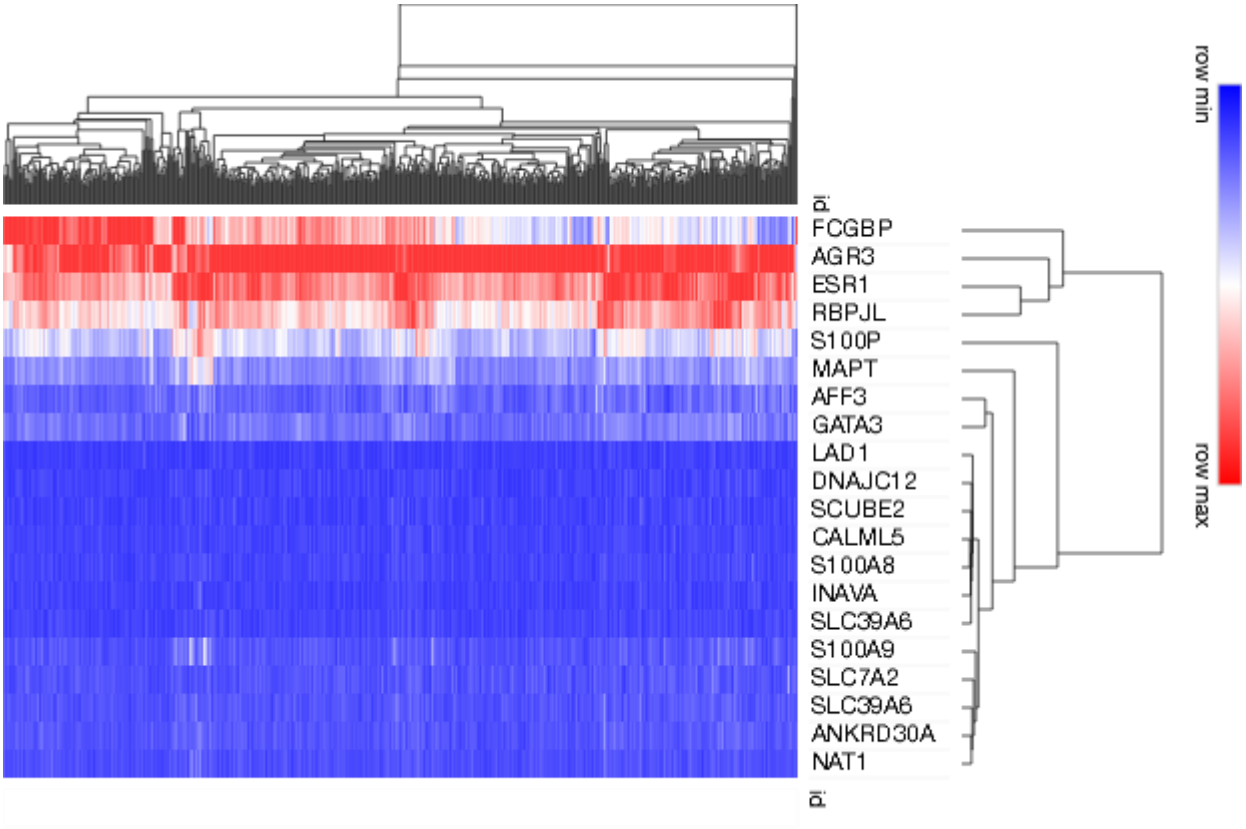


Figure 12. Heat map of Lasso penalty applied on the validation set with a threshold of 4 and 6 years. The above heat map shows, what type of genes act similarly to each other. The color and intensity of the boxes are used to shows changes in gene expression levels, for example, red represents up-regulated genes and green represents down-regulated genes.

Table 2. Go annotation of important genes after applying Lasso penalty on the discovery dataset with the threshold of 4 and 6 years [56] [57].

Illumina ID	Gene Symbol	GO annotation
ILMN_1835913	-	-
ILMN_1792710	DAPK3	Regulates myosin phosphorylation in both smooth muscle and non-muscle cells
ILMN_1801216	S100P	Calcium sensor
ILMN_2310814	MAPT	Promotes microtubule assembly and stability
ILMN_2143566	SLC39A6	Zinc transporter ZIP6
ILMN_1678535	ESR1	Estrogen receptor; Nuclear hormone receptor
ILMN_1728787	AGR3	Encodes a member of the disulfide isomerase (PDI) family of endoplasmic reticulum (ER) proteins
ILMN_1779416	SCUBE2	Signal peptide
ILMN_1796059	ANKRD30A	This gene encodes a DNA-binding transcription factor that is uniquely expressed in mammary epithelium and the testis.
ILMN_2406656	GATA3	Trans-acting T-cell-specific transcription factor GATA-3
ILMN_1750394	SLC39A6	Zinc transporter ZIP6

Table 3. Go annotation of important genes after applying Lasso penalty on the validation dataset with the threshold of 4 and 6 years [57, 56].

Illumina ID	Gene Symbol	GO annotation
ILMN_2302757	FCGBP	Fc fragment of IgG binding protein
ILMN_1835913	-	-
ILMN_1775235	AFF3	Putative transcription activator
ILMN_1782389	LAD1	Anchoring filament protein
ILMN_1651329	-	-
ILMN_1801216	S100P	Calcium sensor and contribute to cellular calcium signaling
ILMN_1749118	CALML5	Calcium-binding protein
ILMN_1785570	-	-
ILMN_2269256	DNAJC12	Associated with complex assembly, protein folding, and export.
ILMN_2310814	MAPT	Microtubule-associated protein tau
ILMN_2143566	SLC39A6	Zinc transporter ZIP6
ILMN_1678535	ESR1	Estrogen receptor
ILMN_1781400	SLC7A2	Cationic amino acid transporter 2
ILMN_1728787	AGR3	Encodes a member of the disulfide isomerase (PDI) family of endoplasmic reticulum (ER) proteins
ILMN_1750974	S100A9	Calcium- and zinc-binding protein
ILMN_1779416	SCUBE2	Signal peptide
ILMN_1796059	ANKRD30A	This gene encodes a DNA-binding transcription factor that is uniquely expressed in mammary epithelium and the testis
ILMN_2406656	GATA3	Trans-acting T-cell-specific transcription factor GATA-3
ILMN_1729801	S100A8	Calcium- and zinc-binding protein
ILMN_1688071	NAT1	This enzyme helps metabolize drugs and other xenobiotics, and functions in folate catabolism
ILMN_1798870	RBPJL	Recombining binding protein suppressor of hairless-like protein
ILMN_1713952	INAVA	Required for optimal pattern recognition receptor (PRR)-induced signaling, cytokine secretion, and bacterial clearance
ILMN_1750394	SLC39A6	Zinc transporter ZIP6

Table 4. Go annotation of important genes after applying Lasso penalty on the discovery dataset with the threshold of 2 and 10 years [56, 57].

Illumina ID	Gene Symbol	GO annotation
ILMN_1651958	MGP	The protein acts as an inhibitor of vascular mineralization and plays a role in the bone organization
ILMN_1722489	TFF1	Stable secretory proteins expressed in gastrointestinal mucosa
ILMN_2365465	XBP1	Functions as a transcription factor during endoplasmic reticulum
ILMN_1678535	ESR1	Estrogen receptor
ILMN_1713086	RPL27A	Ribosomes, the organelles that catalyze protein synthesis
ILMN_1728787	AGR3	Encodes a member of the disulfide isomerase (PDI) family of endoplasmic reticulum (ER) proteins
ILMN_1694742	RPS29	Ribosomes, the organelles that catalyze protein synthesis
ILMN_1809433	XBP1	Functions as a transcription factor during endoplasmic reticulum
ILMN_1654151	COX6C	Catalyzes the electron transfer from reduced cytochrome c to oxygen
ILMN_1688480	CCND1	Phosphorylates and inhibits members of the retinoblastoma
ILMN_1704500	STAP2	Signal transducing adaptor family member 2
ILMN_1788874	SERPINA3	Plasma protease inhibitor
ILMN_1663799	RPL32	This gene encodes a ribosomal protein
ILMN_1730622	EVL	Enhances actin nucleation and polymerization.
ILMN_1661917	-	-
ILMN_1683271	TMSB4X	Binds to and sequesters actin monomers (G actin) and therefore inhibits actin polymerization (44 aa)

Table 5. Go annotation of genes after applying Lasso penalty on the validation dataset with the threshold of 2 and 10 years [57, 56].

Illumina ID	Gene Symbol	GO annotation
ILMN_1775235	AFF3	Function in lymphoid development and oncogenesis.
ILMN_1691884	STC2	Has an anti-hypocalcemic action on calcium and phosphate homeostasis
ILMN_1722489	TFF1	Stable secretory proteins expressed in gastrointestinal mucosa
ILMN_1703891	TBC1D9	May act as a GTPase-activating protein for Rab family protein
ILMN_2365465	XBP1	Functions as a transcription factor during endoplasmic reticulum
ILMN_2269256	DNAJC12	Members of this family of proteins are associated with complex assembly, protein folding, and export
ILMN_1678535	ESR1	Estrogen receptor
ILMN_1774287	CFB	This gene encodes complement factor B
ILMN_1781400	SLC7A2	Functions as permease involved in the transport of the cationic amino acids
ILMN_1728787	AGR3	Encodes a member of the disulfide isomerase (PDI) family of endoplasmic reticulum (ER) proteins
ILMN_1779416	SCUBE2	Signal peptide
ILMN_1688480	CCND1	phosphorylates and inhibits members of the retinoblastoma
ILMN_2406656	GATA3	Trans-acting T-cell-specific transcription factor GATA-3
ILMN_1656791	RPS6	Controlling cell growth and proliferation
ILMN_1788874	SERPINA3	Plasma protease inhibitor
ILMN_1680110	AGAP11; ADIRF	Promotes adipogenic differentiation and stimulates transcription initiation
ILMN_1807825	LY86	Lymphocyte antigen 86
ILMN_1811387	TFF3	Stable secretory proteins expressed in gastrointestinal mucosa
ILMN_1750394	SLC39A6	Zinc transporter ZIP6

Table 6. List of two best genes in both validation and discovery set [56, 57].

Illumina ID	Gene Symbol	GO annotation
ILMN_1678535	ESR1	Estrogen receptor
ILMN_1728787	AGR3	Encodes a member of the disulfide isomerase (PDI) family of endoplasmic reticulum (ER) proteins

4: Discussion

The main goal of this project was to use Machine Learning to predict relapse and overall survival in breast cancer patients using gene expression profiling. To this end, we used different Machine Learning Algorithms such as Logistic Regression, Support Vector Machines, and Random Forest. In addition, for increasing robustness in the prediction of different algorithms all the selected models combined with other robust methods such as Feature Selection methods (Lasso, Recursive Feature Elimination) and also Cross-Validation which we describe as split train and a test set that involves splitting the experimental set into two groups (discovery and validation) multiple times and then retesting them [39].

4.1: Importance of Feature Selection in Machine Learning Systems

What makes Feature Selection crucial and essential in a Machine Learning system is because of directing the system to only select the most useful variables to increase the efficiency of the model[52]. It can do that by minimizing model dimensionality, thus reducing overfitting and also reducting system complexity[52]. Furthermore, Feature Selection can also be useful in optimizing of bias-variance trade-off in a Machine Learning system[52].

The two key parameters in bias and variance are the training set error and the test set error [39]. If the model does very well on the training set, but relatively weak on the test set, that means there is an overfitting problem in the training set, and the model can not generalize well in the Cross-Validation samples [39]. This means that there is a high variance. On the other hand, if the error of the training set is lower than the test set, that means there is a high bias in the system [39]. Since the model does not even do very well on the training set. Therefore, it does not even fit the training data, and that's called underfitting the data, or the algorithm has a high bias [39].

In cases where there is a high error in the training set and even worse in the test set [52]. So, In this case, there is high bias because it is not doing that well on the training set, and high variance in the system [52]. The best example is when there is a low error in the training set and test set which means there is a low bias and low variance in the system [52].

Choosing Lasso as one of the Feature Selection methods helped us to have a bias-variance trade-off and therefore lead to better results in the test set [52]. There is a high variance problem in our data, in other words, the amount of variation in a given variable is high, or there are many noise and irrelevant results that are difficult for the model to handle which lead to high error in the test set and low error in the training set [52]. On the other hand, if there is no variance, the data may be useless. The ability of the Machine Learning model to find a real relationship is called Bias [52]. As a result, there should be a trade-off between bias and variance to obtain the best prediction result, and that's precisely why the Lasso Feature Selection method has been used in this project [52].

4.1.1: Comparison of Lasso and RFE method

Based on the result is shown in Table 1, the best results are when Lasso Logistic Regression combined with the Lasso regularization method with only one percent validation test error.

Although the accuracy of the Logistic Regression decreased from 77 % with RFE to 73 % with Lasso, the validation test result improved dramatically by choosing Lasso instead of RFE.

RFE only considers Feature Selection based on the performance of the model, but Lasso is a better method since it is not a univariate Feature Selection method [39]. Univariate Feature Selection means importance of each feature evaluate individually to determine importance of single feature with the response variable [52]. However, in non-univariate methods such as Lasso, all the features are evaluated together [39]. In other words, the RFE method can measure the linear relationship between each feature and the response variable.

However, the Lasso can select the best features by imposing other features to be close to zero [19]. It is very beneficial for reducing the number of features that are required, yet not necessarily for data interpretation [19]. Therefore, Lasso helped the model to generalize well, and that leads to a better result in the validation test with only a 1 percent error rate. Besides, RFE is not computationally efficient because it repeatedly builds a model, ranks the best and the worst feature and repeats the process until it ranks all the features [39].

4.2: Evaluation of Random Forest, SVM, and RFE

The other model that works best in the validation test is Random Forest. Random Forest is known for overcoming overfitting issues by building a collection of decision trees which makes it much more robust than a single decision tree [21].

Random Forests can consist of one to infinite decision trees, which each tree builds based on a random selection of the observations from the dataset and also a random selection from the available features [23]. Therefore, all the trees will have different sets of features or data points, and this feature of Random Forest assures that the trees are de-correlated and thus less prone to overfitting problems [23]. On the other hand, the application of randomization methods applied both in bagging sample and Feature Selection leads to the selection of uninformative features for node splitting for each tree in Random Forest [23]. As a result, Random Forest will have weak predictions; notably, there is a high dimensional data set. As shown in Table 1, the accuracy of the Random Forest model is less than Logistic Regression and SVM models and also lower than Lasso in the validation test.

SVM is another classification approach that has been used in this project as one of the Feature Selection methods that helps in finding the optimal hyperplane [58]. In the SVM model, there is a hyperparameter C and γ that can be used to choose the best hyperplane with the maximum margin to reduce the overfitting problem [32]. SVM and Lasso both have a hyperparameter to control overfitting problems.

Overall, RFE is not the best method to choose to reduce the number of the feature since after selecting the best features the model will be overfitted, however, the combination of RFE with SVM with its important hyperparameters can reduce overfitting problem. Therefore, the stability of RFE strongly depends on the type of model that is chosen for ranking features at each iteration [39]. Another disadvantage of RFE is that it is very computationally expensive since at each iteration fits the model to remove the weakest feature and also has a high rate of overfitting problem afterward, So instead of that Lasso is the better alternative method that can be used for both Feature Selection and overfitting problem [39].

All the above results are represented when the threshold of 6 years was chosen to separate data into a group of high survival and low survival. In addition to that, a new threshold selected to separate those that lived less than 2 years and died after 10 years and that leads to the highest accuracy of 86%. 8 years gap between the two groups of high and low survival helped the model to learn all the pattern and therefore a better prediction. However, choosing the 6 years threshold will increase the noise of the data set, and consequently, learning from noise will lead to a lower accuracy rate.

4.3: Sources of error in the validation set

In Machine Learning the data set play a critical role in the performance of an algorithm. Therefore, if the data is very noisy or the selected features are not very informative for the models to learn through the data and predict the outcome, then the best models may fail to perform well. In our project, we used the survival rate as our target value. Survival rate indicates the latest status of a patient when they saw their physician for the last time, so the big challenge is that we can not make sure those patients who alive are still alive at the moment. The other issues could be the selection of the best threshold to categorize the data into two groups of high and low survival. The reason to chose the 4 and 6 years threshold was mainly to find the best way not to lose a lot of data and at the same time have almost 50/50 division of high and low survival data into two groups. Our results show that the selection of different thresholds will have a significant effect on the model's performance. even choosing the best threshold may be a temporal solution to predict the survival rate, and therefore, by choosing a different data set, the results may change completely.

4.4: Go annotation

Go annotation provides information about the function of a particular gene and it is made by associating a gene or gene product with a GO term[59]. Therefore, GO annotations represent information about how a gene functions at the molecular level and in what biological processes or pathways are involved [59]. Tables 2, 3, 4 and 5 shows the Go annotation of important genes detected when applying the Lasso penalty with Logistic regression on the Discovery data set and Validation data set. After using different algorithms and Feature Selection techniques, there were two genes, such as ESR1 and AGR3 there were seen in all the results provided above [81]. It shows that these two genes can have an essential role in the Longevity of the patient or classifying the observations into two groups of high and low survival [81].ESR1 gene is responsible for encoding estrogen receptors. Estrogen and its receptors are necessary for sexual development and reproductive function [81]. However, they also play an essential role in tissues like bone. Furthermore, estrogen receptors are also involved in pathological processes such as breast cancer because these receptors get signals from estrogen hormones that could promote abnormal cell growth [81].

AGR3 gene is responsible for proteins that catalyze thiol-disulfide interchange reactions and protein folding [82]. The results of other studies show that breast cancer cells secrete AGR3, and therefore it is normal for AGR3 to overexpressed in the breast cancer cells [82]. Furthermore, it also regulates tumor cell adhesion and migration [82].

The results confirmed that ESR1 and AGR3 are one of the 2 primary genes that overexpressed in breast cancer cells. In addition to AGR3 and ESR1, SLC39A6 was also regularly detected as important genes in breast cancer.SLC39A6 gene product is a zinc transporter ZIP6 which is a protein and it is involved in protein, nucleic acid, carbohydrate, lipid metabolism[60]. Also, it controls the regulation of gene transcription and cell growth[61]. Many studies show zinc transporter ZIP6 regulates intracellular and extracellular Zn levels [61]. Therefore, high and low levels of Zn transporters protein expression showed a direct link to some diseases, such as Alzheimer's disease, diabetes, and cancers [61]. Studies shows, ZIP6 expression levels are down-regulated in high-grade primary breast tumors and it leads to invasion and metastasis [61]. Furthermore, it was found that a high level of ZIP6 protein expression has a direct correlation with a longer relapse-free survival period in patients with breast cancer, and also ZIP6 levels are account as a poor prognostic factor in primary breast tumors in breast cancer patients[61].

Another gene with the name of SCUBE2 was detected as an important gene that is responsible to produce signal peptide. The signal peptide is a short peptide with 16-30 amino acids located at most of N-terminus newly synthesized proteins and its function is to direct a cell to translocate the protein, usually to the cellular membrane[62]. Studies shows, upregulation of SCUBE2 expression may enhance metastasis for patients with breast cancer [63].

GATA-3 was also detected as another important gene in breast cancer. GATA-3 is a transcription factor that controls human growth and differentiation[64]. Gene expression profiling in several studies shown that GATA-3 in the Luminal A subtype of breast cancer is highly expressed [64].

4.5: Comparison of other Machine Learning work in breast cancer

There are different studies available that applied different Machine Learning approaches for identifying gene biomarkers for the treatment of breast cancer. The most recent studies show the use of a supervised learning model to estimate which breast cancer patients will survive more than 5 years after undertaking a specific treatment therapy [15]. So the goal was to build a model based on the combination of the treatment and survivability of the patient [15].

Different classes were identified based on different treatments such as surgery, hormone therapy, and radiotherapy with a patient status as living or dead[15]. Their gene expression data set consisted of 24,368 genes for each of the 347 samples[15]. Because of the large number of features and overfitting problems they performed different filter features selection methods such as chi_square and information gain. Besides, another feature method such as utilizing minimum redundancy maximum relevance (mRMR) was also applied to the remaining features to select those attributes having the highest relevance in productivity of the model[15]. Therefore, it can be used to ranks features according to the minimal-redundancy-maximal-relevance criterion. In this algorithm, relevance is calculated by using the F-statistic and redundancy measured by Pearson correlation[15].

After the selection of the best features, different classifiers, such as naive Bayes, random forest, and SVM was used in that study [15]. These models used a one-versus-rest scheme called the multiclass problem. This approach includes classifying one class against the remaining classes and then removing that class from the dataset [15]. Repeat the previous step for another class and so on.

Another study from BMC Bioinformatics in 2017 shows utilizing the minimum redundancy maximum relevance approach with K-nearest neighbors (KNN), Naive Bayes classifier (NB), Support vector machine (SVM), and Random Forest classifiers demonstrate better accuracy results[65]. As a result, they also used naive Bayes, random forest classifiers to power classification than other models[65]. Furthermore, the classification model was combined using a 10-fold cross-validation technique to calculate a more accurate estimate of model accuracy [65].

Finally the performance of models evaluated by calculating accuracy, sensitivity, F1-measure, and specificity[65]. There is some main difference between Lasso method and mRMR algorithm. Lasso removes redundant features based on their coefficients[66]. So to select the best feature, reduce the weights of non-relevance features to near zero[66]. However, this approach is fitting or link to the embedded feature selection techniques, which means a different subset of features is made into the classifier construction instead of the filter type feature selection technique[66].

In filter type feature selection methods the algorithm chooses those features that are highly predictive but at the same time uncorrelated to other features that are independent of any machine learning algorithms[67]. So, the best features are selected based on their statistical tests score such as Pearson's Correlation and Chi-Square that indicated their correlation with the response or outcome variable[67]. An example of a filter-based feature selection is the Maximum Relevance Minimum Redundancy (mRMR) algorithm[67].

On the other hand, embedded feature selection method such as Lasso are more likely to wrapper methods such as recursive feature elimination since they both optimizing the performance of a learning algorithm, the only difference is, in RFE goal is to find the best feature subset based on the classifier performance but in Lasso objective is to add penalty against to reduce overfitting or variance of a model[68].

In our study, we studied the use of different machine learning algorithms and also with combination by different feature selection approach to predict the likelihood of breast cancer survival. Many studies believe genomic data will help us to better predict the survival rate of these patients and therefore will provide a better and more personalized treatment option. The major challenge in survival prediction models is coping with a large number of dimensions of gene expression data. Many studies used different dimensionality reduction techniques to overcome this problem in different ways.

Many different Machine Learning methods and different dimensionality reduction techniques have already been applied for microarray data analysis such as k-nearest neighbors, Hierarchical Clustering, Support Vector Machines, and Bayesian networks[66]. Most of the studies show the embedded model such as Random Forest and deep learning-based multi-model ensemble method gives a much better prediction result or relationships among the features[66].

As suggested in the literature[69], the Deep Learning-based multi-model ensemble method was one of the most accurate methods which are made by two stages. In the first stage, multiple different classifications stacked together to provide better performance rather than the individual model, and then the prediction of the collection of the model is used as an input to the second stage learning model[69]. The second sage is trained based on the combination of prediction for further optimization and forming the final set of predictions[69]. So it can automatically learn non-linear relationships from the data and make improve the accuracy of the model prediction[69].

In addition to Random Forest and Deep Learning-based multi-model ensemble models, it was found that kNN is one of the simplest classifications that can be used in cancer classification prediction[66].SVMs also considered an effective model for cancer classification. However, for SVMs is challenging to choose the best kernel for specific issues[66]. In addition to the SVM and kNN model, Decision Tree is one of the widely used models in this field but the major problem with this model is that it over-fits the model and does not generalize well in the test set[66].

4.6: Future work

Deep Learning is an artificial intelligence function that inspired by the human brain, which can process data and detect patterns for decision-making purposes [83]. Deep Learning is a subgroup of Machine Learning in artificial intelligence. In Deep Learning, neural networks are used to learn from unstructured or unlabeled data. A neural network is made up of many cells or neurons that are connected to provide the desired result or output [84].

I think a Deep Learning model is also worth to try since we can have less worry about the Feature Selection part because the classifier prediction more depends on the dataset and its general complexity [84]. Also, Deep Learning algorithms don't perform well with a small data set, and that's because Deep Learning algorithms need a large amount of data to understand the pattern correctly [84]. So It is not feasible to only use about 500 amounts of data with Deep Learning models. For that reason, we may need to collect more breast cancer data to what we have in future research. Deep Learning using multiple layers and numerous parameters to extract essential features from the input data[84]. In Deep Learning, several transformations occur because there are several layers between input and output layers in the neural network to combine layers and layers of features [84]. However, if we compare it with the SVM model, there usually is only a single transformation [83]. In neural networks choosing the number of layers is an arbitrary choice, and therefore using more layers allows us to extract more information from the data. However, at the same time, too many layers make neural networks more prone to overfitting [83]. In addition to the above-suggested approach, I think searching about what clinical features can add more value to the data set is also very critical.

We could also apply different regression Machine Learning by considering the survival rate column as a continuous variable instead of a binary variable. Random Forest is one of the possible regression models that can be used. Since Random Forest it's great on non-linear relationships and complex learning and it's very easy for the model to understand and interpret the decision boundaries is made in the training set[24]. On the other hand, SVM may not be an ideal regression model because having a large number of features increases the complexity of the model by transforming all the samples into a higher dimension.

Polynomial Regression may be the best model instead of simple linear regression. The Polynomial Regression hypothesis fits the data with a curvilinear relationship between the output variable and the independent variables[70]. So the big difference of Polynomial with linear regression is that there is a parameter in the former model that helps the model to choose a different degree of polynomial based on the target variable and the predictor[70]. As an example, if we choose a 1-degree polynomial that means the hypothesis will be a simple linear regression so choosing a higher degree of the polynomial will increase the complexity of the model to fit the data well rather than linearly[70].

5: References

1. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2019. *CA. Cancer J. Clin.* **69**, 7–34 (2019).

2. La Vecchia, C., Negri, E. & Boyle, P. Reproductive factors and breast cancer: An overview. *Sozial- und Präventivmedizin SPM* **34**, 101–107 (1989).

3. Kinsella, M. D., Nassar, A., Siddiqui, M. T. & Cohen, C. Estrogen receptor (ER), progesterone receptor (PR), and HER2 expression pre- and post- neoadjuvant chemotherapy in primary breast carcinoma: A single institutional experience. *Int. J. Clin. Exp. Pathol.* **5**, 530–536 (2012).

4. DeSantis, C., Ma, J., Bryan, L. & Jemal, A. Breast cancer statistics, 2013. *CA. Cancer J. Clin.* **64**, 52–62 (2014).

5. Jemal, A., Thomas, A., Murray, T. & Thun, M. Estimated New Cancer Cases. *Cancer Stat.* **52**, 23–47 (2002).

6. Carey, L. A. *et al.* Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study. *J. Am. Med. Assoc.* **295**, 2492–2502 (2006).

7. Onitilo, A. A., Engel, J. M., Greenlee, R. T. & Mukesh, B. N. Breast Cancer Subtypes Based on ER/PR and Her2 Expression: Comparison of Clinicopathologic Features and Survival. *Clin. Med. Res.* **7**, 4–13 (2009).

8. Iqbal, B. & Buch, A. Hormone receptor (ER, PR, HER2/neu) status and proliferation index marker (Ki-67) in breast cancers: Their onco-pathological correlation, shortcomings and future trends. *Med. J. Dr. D.Y. Patil Univ.* **9**, 674 (2016).

9. Callahan, R. & Hurvitz, S. HER2-Positive Breast Cancer: Current Management of Early, Advanced, and Recurrent Disease. *Curr Opin Obs. Gynecol* **29**, 997–1003 (2012).

10. Shi, W., Oshlack, A. & Smyth, G. K. Optimizing the noise versus bias trade-off for Illumina whole genome expression BeadChips. *Nucleic Acids Res.* **38**, (2010).

11. Price, M. E. *et al.* Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. *Epigenetics and Chromatin* **6**, 1–15 (2013).

12. Varoquaux, G. *et al.* Scikit-learn. *GetMobile Mob. Comput. Commun.* **19**, 29–33 (2015).

13. Schmid, R. *et al.* Comparison of normalization methods for Illumina BeadChip HumanHT-12 v3. *BMC Genomics* **11**, (2010).

14. Ozsolak, F. & Milos, P. M. RNA sequencing: Advances, challenges and opportunities. *Nat. Rev. Genet.* **12**, 87–98 (2011).

15. Tabl, A. A., Alkhateeb, A., ElMaraghy, W., Rueda, L. & Ngom, A. A machine learning approach for identifying gene biomarkers guiding the treatment of breast cancer. *Front. Genet.* **10**, 1–13 (2019).

16. Jordan, M. I. & Mitchell, T. M. Machine learning: Trends, perspectives, and prospects. *Science (80-.).* **349**, (2015).

17. Turner, J. & Charniak, E. Supervised and unsupervised learning for sentence compression. *ACL-05 - 43rd Annu. Meet. Assoc. Comput. Linguist. Proc. Conf.* 290–297 (2005) doi:10.3115/1219840.1219876.

18. Machine Learning: Supervised Vs Unsupervised Learning – lakshaysuri. https://lakshaysuri.wordpress.com/2017/03/19/machine-learning-supervised-vs-unsupervised-learning/.

19. Meier, L., Van De Geer, S. & Bühlmann, P. The group lasso for logistic regression. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **70**, 53–71 (2008).

20. Japkowicz, N. & Shah, M. Evaluating Learning Algorithms. *Eval. Learn. Algorithms* (2011) doi:10.1017/cbo9780511921803.

21. Qi, Y. Random forest for bioinformatics. *Ensemble Mach. Learn. Methods Appl.* 307–323 (2012) doi:10.1007/9781441993267_10.

22. Chen, X. & Ishwaran, H. Random forests for genomic data analysis. *Genomics* vol. 99 323–329 (2012).

23. Liaw, A. & Wiener, M. Classification and Regression by randomForest. *Merck Res. Lab.* (2014).

24. Random Forest Simple Explanation - Will Koehrsen - Medium. https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d.

25. Kuh, A. & De Wilde, P. Comments on 'Pruning error minimization in least squares support vector machines'. *IEEE Trans. Neural Networks* **18**, 606–609 (2007).

26. Noble, W. S. What is a support vector machine? *Nat. Biotechnol.* **24**, 1565–1567 (2006).

27. Jakkula, V. Tutorial on Support Vector Machine (SVM). *Sch. EECS, Washingt. State Univ.* 1–13 (2011).

28. General classification hyperplane representation of SVM algorithm. | Download Scientific Diagram. https://www.researchgate.net/figure/General-classification-hyperplane-representation-of-SVM-algorithm_fig5_330557084.

29. Machine Learning Basics: Support Vector Machines - Data Driven Investor - Medium. https://medium.com/datadriveninvestor/machine-learning-basics-support-vector-machines-358235afb523.

30. Ladwani, V. M. Support vector machines and applications. *Comput. Vis. Concepts, Methodol. Tools, Appl.* 1381–1390 (2018) doi:10.4018/978-1-5225-5204-8.ch057.

31. Schölkopf, B. Slides- Learning with kernels. *J. Electrochem. Soc.* **129**, 2865 (2002).

32. Huang, X., Maier, A., Hornegger, J. & Suykens, J. A. K. Indefinite kernels in least squares support vector machines and principal component analysis. *Appl. Comput. Harmon. Anal.* **43**, 162–172 (2017).

33. Kurt, I., Ture, M. & Kurum, A. T. Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. *Expert Syst. Appl.* **34**, 366–374 (2008).

34. Day, W. H. E. & Edelsbrunner, H. Efficient algorithms for agglomerative hierarchical clustering methods. *J. Classif.* **1**, 7–24 (1984).

35. Fürnkranz, J. & Flach, P. A. An Analysis of Rule Evaluation Metrics. *Proceedings, Twent. Int. Conf. Mach. Learn.* **1**, 202–209 (2003).

36. Cross-validation (statistics) - Wikipedia. https://en.wikipedia.org/wiki/Cross-validation_(statistics).

37. Anguita, D., Ghelardoni, L., Ghio, A., Oneto, L. & Ridella, S. The 'K' in K-fold cross validation. *ESANN 2012 proceedings, 20th Eur. Symp. Artif. Neural Networks, Comput. Intell. Mach. Learn.* 441–446 (2012).

38. Rodríguez, J. D., Pérez, A. & Lozano, J. A. Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **32**, 569–575 (2010).

39. Yan, K. & Zhang, D. Feature selection and analysis on correlated gas sensor data with recursive feature elimination. *Sensors Actuators, B Chem.* **212**, 353–363 (2015).

40. Centre, G. C. a Genetic Algorithm Based Wrapper Feature Selection Method for Classification of Hyperspectral Images. *Archives* 397–402 (2006).

41. Curtis, C., Shah, S. P., Chin, S. & Turashvili, G. Europe PMC Funders Group The genomic and transcriptomic architecture of 2 , 000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).

42. Peng, C. Y. J., Lee, K. L. & Ingersoll, G. M. An introduction to logistic regression analysis and reporting. *J. Educ. Res.* **96**, 3–14 (2002).

43. Lee, W. S. & Liu, B. Learning with Positive and Unlabeled Examples Using Weighted Logistic Regression. *Proceedings, Twent. Int. Conf. Mach. Learn.* **1**, 448–455 (2003).

44. Bewick, V., Cheek, L. & Ball, J. Statistics review 14: Logistic regression. *Crit. Care* **9**, 112–118 (2005).

45. Coppersmith, D., Hong, S. E. J. & Hosking, J. R. M. Partitioning nominal attributes in decision trees. *Data Min. Knowl. Discov.* **3**, 197–217 (1999).

46. Stamate, D., Alghamdi, W., Stahl, D., Logofatu, D. & Zamyatin, A. PIDT: A novel decision tree algorithm based on parameterised impurities and statistical pruning approaches. *IFIP Adv. Inf. Commun. Technol.* **519**, 273–284 (2018).

47. Raileanu, L. E. & Stoffel, K. Theoretical comparison between the Gini Index and Information Gain criteria. *Ann. Math. Artif. Intell.* **41**, 77–93 (2004).

48. Xia, F., Zhang, W., Li, F. & Yang, Y. Ranking with decision tree. *Knowl Inf Syst* (2007) doi:10.1007/s10115-007-0118-y.

49. Probst, P., Wright, M. N. & Boulesteix, A. L. Hyperparameters and tuning strategies for random forest. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **9**, 1–19 (2019).

50. Bernard, S., Heutte, L. & Adam, S. Influence of hyperparameters on random forest accuracy. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* **5519 LNCS**, 171–180 (2009).

51. machine learning - What is the influence of C in SVMs with linear kernel? - Cross Validated. https://stats.stackexchange.com/questions/31066/what-is-the-influence-of-c-in-svms-with-linear-kernel.

52. Dash, M. & Liu, H. Feature selection for classification. *Intell. Data Anal.* **1**, 131–156 (1997).

53. Ramírez-Hernández, J. A. & Fernandez, E. Control of a re-entrant line manufacturing model with a reinforcement learning approach. *Proc. - 6th Int. Conf. Mach. Learn. Appl. ICMLA 2007* 330–335 (2007) doi:10.1109/ICMLA.2007.35.

54. Powell, A., Bates, D., van Wyk, C. & Darren de Abreu, A. A cross-comparison of feature selection algorithms on multiple cyber security data-sets. *CEUR Workshop Proc.* **2540**, 196–207 (2019).

55. Zhao, S., Mao, X., Lin, H., Yin, H. & Xu, P. Machine Learning Prediction for 50 Anti-Cancer Food Molecules from 968 Anti-Cancer Drugs. *Int. J. Intell. Sci.* **10**, 1–8 (2020).

56. STRING: functional protein association networks. https://string-db.org/.

57. bioDBnet - Biological Database Network. https://biodbnet-abcc.ncifcrf.gov/db/db2db.php.

58. Frauke, F. & Christian, I. Evolutionary Tuning of Multiple SVM Parameters. *Neurocomputing* **64**, 107–117 (2005).

59. Camon, E. The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res.* **32**, 262D – 266 (2004).

60. Lian, J. *et al.* miR-192, a prognostic indicator, targets the SLC39A6/SNAIL pathway to reduce tumor metastasis in human hepatocellular carcinoma. *Oncotarget* **7**, 2672–2683 (2016).

61. Takatani-Nakase, T. Zinc transporters and the progression of breast cancers. *Biological and Pharmaceutical Bulletin* vol. 41 1517–1522 (2018).

62. Lin, Y. C., Chen, C. C., Cheng, C. J. & Yang, R. B. Domain and functional analysis of a novel breast tumor suppressor protein, SCUBE2. *J. Biol. Chem.* **286**, 27039–27047 (2011).

63. Lin, Y. C., Lee, Y. C., Li, L. H., Cheng, C. J. & Yang, R. B. Tumor suppressor SCUBE2 inhibits breast-cancer cell migration and invasion through the reversal of epithelial-mesenchymal transition. *J. Cell Sci.* **127**, 85–100 (2014).

64. Voduc, D., Cheang, M. & Nielsen, T. GATA-3 expression in breast cancer has a strong association with estrogen receptor but lacks independent prognostic value. *Cancer Epidemiol. Biomarkers Prev.* **17**, 365–373 (2008).

65. Radovic, M., Ghalwash, M., Filipovic, N. & Obradovic, Z. Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. *BMC Bioinformatics* 1–14 (2017) doi:10.1186/s12859-016-1423-9.

66. Yuvarajan, V., Sathiyabhama, B. & Udhaya Kumar, S. A Comparison of Machine Learning Techniques for Survival Prediction in Breast Cancer Gene Expression Data. *SSRN Electron. J.* 1–8 (2018) doi:10.2139/ssrn.3126112.

67. Lanzi, P. L. Fast feature selection with genetic algorithms: A filter approach. in *Proceedings of the IEEE Conference on Evolutionary Computation, ICEC* 537–540 (1997). doi:10.1109/icec.1997.592369.

68. Karegowda, A. G., Manjunath, A. S. & Jayaram, M. A. Feature Subset Selection Problem using Wrapper Approach in Supervised Learning. *Int. J. Comput. Appl.* **1**, 13–17 (2010).

69. Xiao, Y., Wu, J., Lin, Z. & Zhao, X. A deep learning-based multi-model ensemble method for cancer prediction. *Comput. Methods Programs Biomed.* **153**, 1–9 (2018).

70. Ostertagová, E. Modelling using polynomial regression. in *Procedia Engineering* vol. 48 500–506 (2012).