Machine Learning to Predict Breast Cancer Aggressiveness Based on Immunohistochemistry HER2 Test

Sami Ali Choudhry
Computer Science
University of Windsor
Windsor, Canada
choud116@uwindsor.ca

Ali Hassan
Computer Science
University of Windsor
Windsor, Canada
hassa135@uwindosr.ca

Balsharan Singh Bedi
Computer Science
University of Windsor
Windosr, Canada
bedib@uwindsor.ca

***Abstract*—Breast Cancer is the most common type of cancer that effects women. Recent advancements in genomics have enabled high resolution tracing of CNA (Copy Number Abnormality) values. These values give important information about various genes that may have a role in causing the cancer. Immunohistochemistry (IHC) HER2 test is used to determine how aggressive the cancer is and also helps determine if and when the patient might need an operation. Using the CNA values in combination with the HER2 test values of the patients in the dataset a classifier is trained to predict the outcome of a HER2 test before conducting the test. This will help save time it takes to actually conduct the detailed test and take necessary precautions before the actual test results are returned from lab.**

***Index Terms*—SVM, Breast Cancer, Dataset Filtering, Classification, Sampling, KPCA, CNA**

I. Introduction

A. What is Breast Cancer?

Whenever there is an enormous amount of growth in the cells of body and began to spread out in different portions of a body there are chances that a patient has cancer and if it is present in breast it is called breast cancer. A breast consists of lobes and ducts each of which has approximately 15 lobes which are further divided into lobules which consists of tiny ducts which helps in the milk generation and thus thin tubes connect lobes, lobules and bulbs by ducts. It also consists of lymph vessels along with blood vessels. The main function of lymph vessel is to carry lymph (colorless, watery fluid) and blood vessels carry blood throughout. The most common types of breast cancer is ductal carcinoma, which starts in the cells of the ducts then there is lobular carcinoma which is found as a common trait in almost all the types of breast cancer also the least occurring one is the inflammatory breast cancer in which it leaves breast red, warm and swollen.

B. What causes Breast Cancer?

One of the most important age range which are at higher risk of breast cancer are those below 45 especially if there are any changes in BRCA1 and BRCA2 genes or if there are any close relatives with such kind of changes in the genes. Also those with higher density near the mammogram are at a greater risk of breast cancer. Tumor Suppressor Genes synthesize a protein called a tumor suppressor protein that helps control cell growth so whenever mutations occur in tumor suppressor genes it may lead to cancer. Some of other factors which leads to breast cancer below

- **Exposure to radiation:** Any previous history of treatment which had exposures to radiations also increases the chances for breast cancer.
- **Early age menstruation:** Because the body is exposed to estrogen for longer period of time it thereby increases the chances for breast cancer.
- **Older age menstruation:** Because the body is exposed to estrogen for longer period of time it thereby increases the chances for breast cancer.
- **Alcohol consumption:** Alcohol increases the level of estrogen in the body.

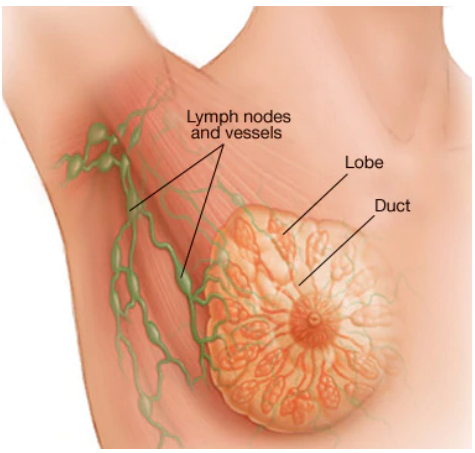

Fig. 1. Breast Structure

Inherited Gene Mutations: Breast cancer can be caused by the genes passed from parent to child. Genes are defined as short segments of DNA (deoxyribonucleic acid) that are present in chromosomes. Our human body is made up of numerous cells and proteins determine the functioning of these cells. DNA are responsible for creating the proteins in our body. We can imagine genes as the instruction provider responsible for cell growth. If there is a wrong instruction provided by a gene, the cell growth will be faulty as well. In other words, we can say that if a gene is erroneous, then this error will be contained in all the cells of that gene. Generally, we can say that changes in DNA can be of 2 types. The first type is those that are inherited from parents and the other type is those that happen as the time passes. The first type or the inherited DNA change is called as mutations and the second type in which DNA changes happen over a period of time is called somatic alterations. Breast cancer can occur when mutation occurs in 2 genes: BRCA1 (BReast CAncer gene one) and BRCA2 (BReast CAncer gene two). The major responsibility of BRCA genes is prevent cell damage and allow normal cell growth in our body. The risk of breast cancer increases when the mutations in these genes are passed from parent to children. But it is also not completely proven that mutation in BRCA1 and BRCA2 can definitely cause breast cancer. There can be some other mutations that can lead to breast cancer as well. For instance, SNPs (single nucleotide polymorphisms) may be linked to high rate of breast cancer in women. On contrary, it is also seen that mutation of BRCA 1 and BRCA2 can lead to breast cancer or ovarian cancer.
According to a research, from all the total women population, 12 % women can have breast cancer. Out of these, women having BRCA1 and BRCA2 mutation have a 7% chance of having breast cancer. In women, when both the mutations take place, breast cancer can occur at a very young age as well as in both the breasts. Men who have BRCA1 mutation are more prone to prostate cancer. Men having BRCA 2 mutation are several times more likely to develop prostate cancer than the men who do not have such a mutation. Also, in men if BRCA1 and BRCA2 mutation takes place, the all other types of cancer can also take place.
Acquired Gene Mutations: Although breast cancer mostly occurs during the mutation of BRCA1 and BRCA2 but it can also occur due to mutations in other genes as well.

- **ATM:** Damaged DNA is repaired by ATM, genetic information is present in the DNA out of which two mutated copies are the reason to cause disease ataxia-telangiectasia, a rare disease that affects brain development thus inheriting one such mutation of gene increases the risk of breast cancer.

- **BARD-1:** BRCA1 gene works with BARD1 to repair DNA damage, mutation within itself thereby increases the risk of breast cancer in women

- **BRIP1:** It also works to repair DNA, thus inheriting one of the mutation gene thereby increases the risk of breast cancer

- **CDH1:** It helps in building up of the protein that binds cells together to form tissue, thus mutation of this also increases the rate of breast cancer, thus the lifetime risk of it is about 83

- **CHEK2:** It helps in building up of the protein that stops the tumour growth, it can atleast double the risk of breast cancer and colon cancer also increases the risk of prostate cancer.

C. Output Feature Selection

- **IHC by HER2:** Immunohistochemistry test result that denotes the aggressiveness of the cancer.

- **ONCO-Tree Code:** Code for the type of tumor and tissue type.

II. Project Overview

The aim of the project is to show best results upon processing the given carcinogenic breast cancer dataset specifically called as Breast Invasive Carcinoma. Initially the given dataset consist of a large amount of unwanted data and thus we begin preprocessing by arranging the data in particular order and deleting the unwanted NaN values. Furthermore, we select the features from the dataset which help us in inferring valuable conclusions while clustering and classification. This process is called data filtering. After this step, sampling is applied across the dataset. This technique is used to populate the dataset with some more entries for which there are less number of samples for that output class in y. We also perform feature reduction by using only the best K features and applying Kernel PCA using RBF as the main kernel. Later, we perform classification by applying Support Vector Machines[1] (linear and rbf kernel), Naive Bayes, Random Forest, and K Nearest Neighbour algorithms. We choose IHCHER2 as the output class for multi-class classification. IHC, or ImmunoHistoChemistry, is a process that is performed on the breast cancer tissue. The IHC plays a vital role in planning the treatment for breast cancer. The IHC is used to see the HER2 receptor protein in the breast cancer tumour. The HER2 score is between 0 to 3+. If this value is between 0 to 1+, the this is HER2 negative. If the score is 3+, the it is HER2 positive. The value between these is called borderline. The performance analysis of each algorithm used for classification will be done. Moreover, the gene data given for patients is clustered so that patients having the similar genes can be clustered together.

III. Problem Statement

Out of all the types of breast cancers Ductal Invasive Carcinoma is the most common. So, we have focused our research on Breast invasive carcinoma. The dataset we have selected is Breast Invasive Carcinoma (TCGA, Cell 2015). The distribution of various types of cancer in the dataset is

given below

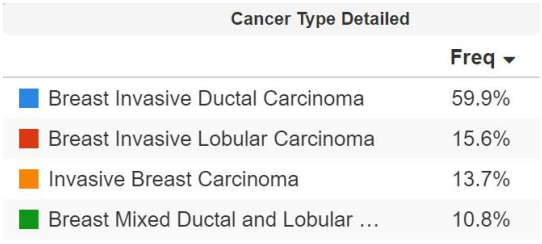

Fig. 2. Types of Cancer in Dataset

The gene data is the (CNA) Copy Number Abnormality also known as Copy Number Variation (CNV). Recent advancements in bio genomics new methods that are capable of extremely high genomic resolution and as a result can trace adetailed amount of copy number variations in the genome. The output file is the Data Clinical Sample which contains the data of all the sample obtained from the patients throughout the period under study.*We aim to predict HER2 test result and ONCO Tree Code for the type of cancer.* The dataset used for the processing of carcinogenic breast cancer are of two types out of which one is data clinical sample which is used for output feature selection and the other one is data linear CNA which consists of CNA about 22247 genes for 817 patients. The HER2 plays a critical role in the surgery of breast cancer. When there are a large number of HER2 receptors, then the cancer cells may receive numerous signals. This receiving tend them to grow and divide. Inaccurate HER2 predictions may prevent women to get the best care possible. So, we aim to train a classifier that can predict the HER2 result based on the gene data of the patient provided.

IV. Techniques

A. Data Preprocessing

The CNA file consists of copy numbers for 22247 genes of 817 patient samples. First, we remove any extra rows that consist of extra information that is irrelevant to the algorithms such as Hugo Symbol and Gene-ID. The data initially consists of negative and positive values. This results in error several feature selection and feature reduction algorithms. Hence, we first scale the data using min/max scaler to bring the CNA values of the features between 0 and 1. The resulting data allows for more algorithms to be used in determining the most optimal features of the dataset. The Data Clinical Sample file consists of output values observed from the clinical samples. This file consists of several features that we can use as Y values. Due to the importance of HER2 we have selected it as the output value as it is an important factor in determining how aggressive the cancer's behavior might be. The column of the output value is extracted from this file and concatenated with the transpose of the gene data file. So, each Y value is concatenated In front of the respective patient's data. The HER2 column has 4 possible values *"Positive"*, *"Negative"*, *"Equivocal"* and *"[Not Available]"*. We filter out the patients that have the value as *"[Not Available]"* later.

B. Feature Extraction

Breast Invasive Carcinoma dataset consists of 22247 features. But, when we are performing classification, not all the features are required. There are some unimportant features that can introduce noise in our data and can reduce the accuracy while classification. There is also one more issue when such high dimension data is used. The issue is that when such high number of features are used to train the model, the time complexity during training becomes very high. This is not good as it leads to poor user experience. The solution to these problems is to choose the important features that can give good results during classification. By taking the above step, our machine learning model will be trained in a shorter period of time and the results will be more accurate. There are many models in machine learning available for feature selection and reduction. We have used SelectKBest to choose the best features from the dataset provided. Also, Principal Component Analysis (PCA) and Kernel Principal Component Analysis (K-PCA) have been used for feature reduction.

- **Select K Best:** The SelectKBest is a technique in which we choose only those features from the dataset that can lead to best predictions during classification. These features selected are those that can reduce the noise in data and help us in reaching a better result. In short, this method is used to select the best predictors for the output class prediction. This technique takes X(training input) and Y(training output) as the parameters. It then returns an array of scores for every feature in the dataset. It then chooses the first K features with the highest scores. Let us say that we choose chi2 as the score function in this algorithm. It then calculates this score between every feature in our dataset with the output value. The value of this score can tell if the feature is important or not. If the score has very small value, then it means that the feature and the output value are not related to each other. A high value of this score means that there is a strong correlation between the output value and the feature. As a result, K such features with strong correlation is selected to improve the accuracy of classification.

- **ANOVA F-value:** This is another technique that can be applied for feature selection. This techniques performs a test in which False Positive Rate is determined. This is also called as the FPR test. The main aim of this test is to control the total amount of False detections made during the FPR test. This method takes the score function as the argument. This function further requires the X, the input dataset, and Y, the class labels as the output. This technique then returns the arrays containing all the scores for each feature and the pvalues. Also, another argument

is the alpha that determines the highest pvalues for the features.

C. Data Sampling

For machine learning algorithm to perform good, both in terms of time and accuracy, the data must not have many dimensions and it must be balanced. We have solved the first problem in the previous section. But we need to solve the second problem of data balancing. By balancing we mean that there must be equal number of samples of each class while the classification algorithm is getting trained. In our dataset, the data is unbalanced. So, for some output classes there are more samples than the samples belonging to other class. So, data balancing is vital to improve the accuracy while classification. We have used SMOTE sampling for balancing our dataset.

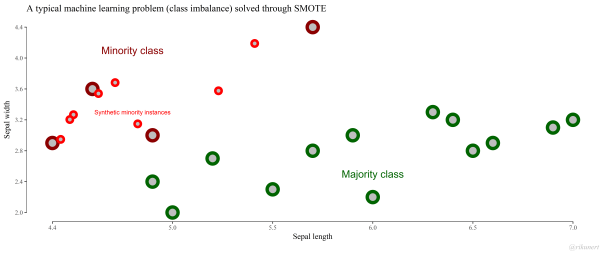

Fig. 3. SMOTE Re-sampling

As seen in the figure above, there are more samples of one class as compared to the other. In such a situation where one class dominates the other, the machine learning algorithm performs very poorly. Now, to solve this problem more points belonging to the red class need to be added. So, SMOTE firstly draws line between the samples belonging to the minority class. It then assumes the new points on these lines. This process can be shown below in the figure.

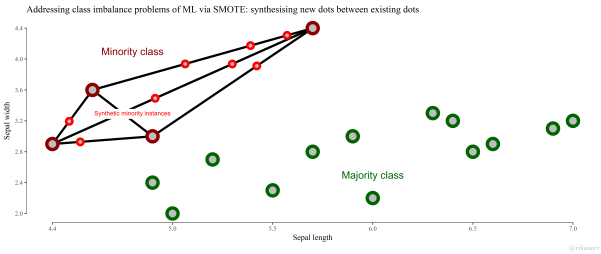

Fig. 4. SMOTE Execution

Now we will apply SMOTE sampling on Breast Invasive Carcinoma dataset. Initially the dataset looks as seen below,

The above figure represents the dataset without applying sampling. As we can see, the count of samples belonging to the 'Negative' class is the highest. For the 'Equivocal', 'Not Available', 'Positive' classes the number of samples are less. The total number of samples for 'Intermediate' class are very less. So, we will remove the samples belonging to this class. Next, we will apply sampling on the points belonging to the 'Equivocal', 'Not Available', 'Positive' classes. The output

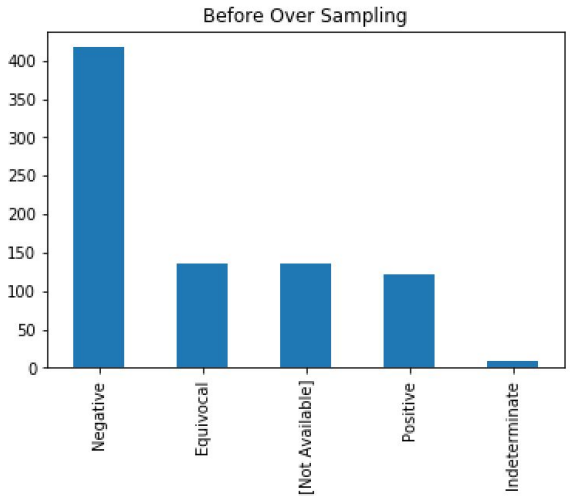

Fig. 5. Raw Dataset

after applying SMOTE sampling can be seen below in the figure.

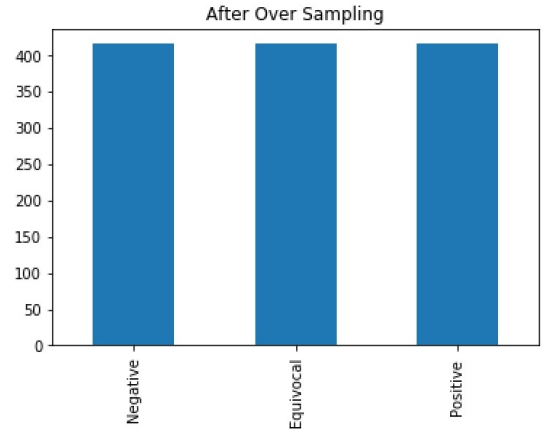

Fig. 6. Over sampled Dataset

D. Feature Reduction

We have used two approaches for feature reduction, Principal Component Analysis (PCA) and Kernel Principal Component Analysis (K-PCA).

- **PCA:**
  Principal Component Analysis(PCA) is a linear dimensionality reduction technique. The data is represented in a lower dimension such that the variance or the correlation matrix of this data in this new dimension is maximized. In the first step the eigen vectors of the correlation matrix are computed. The eigenvectors having the maximum eigen values are chosen. Thes values can be used to find the variance in the original dataset. Moreover, these eigen values represent the behaviour of the system. Finally, the

data is represented in a lower dimension but with some data loss and by retaining the variance of the original dataset. As we can see below in the figure, the dataset has 2 feature values represented on axis by variable 1 and variable 2. PCA will represent the data using the linear line shown by PCA 1 and PCA 2. As the aim to maximize the variance, so the line represented by PCA 1 is chosen.

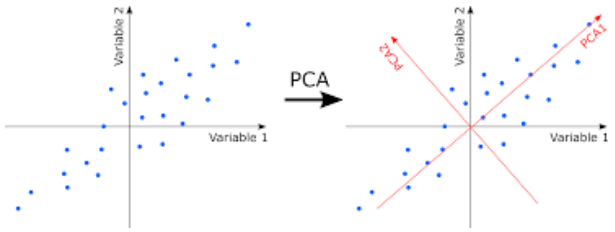

Fig. 7. Principal Component Analysis

- **K-PCA:**
  Kernel PCA computes the Gram matrix and then it finds the m no. of eignevectors, eigen values of this matrix. It then finds the m projections of each dimension of x. We have used the rbf kernel that uses the Gaussian similarity function given by.

$$k(x,y)=\exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$$

Fig. 8. Radial Basis Function

Kernel PCA uses the kernel trick to find a principal component. When the dimension is reduced using kernel PCA, the points belonging to 2 different classes are well separated on this 1-dimension. The figure below shows the circles0.3 dataset. This dataset contains only 2 features. We now apply the Kernel PCA on the below dataset.

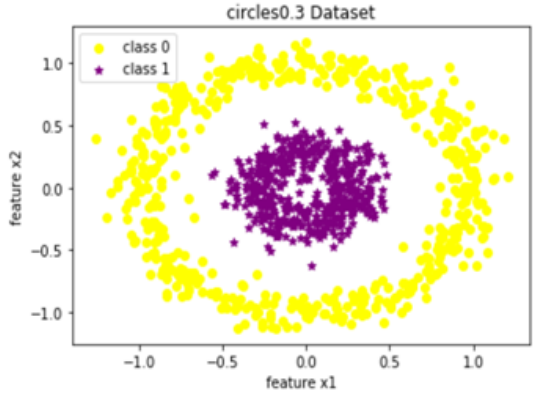

Fig. 9. Circles0.3 Dataset in 2D

E. Description of Solution

a) Classification: Classification is a supervised learning technique that takes a subset of the dataset as training data

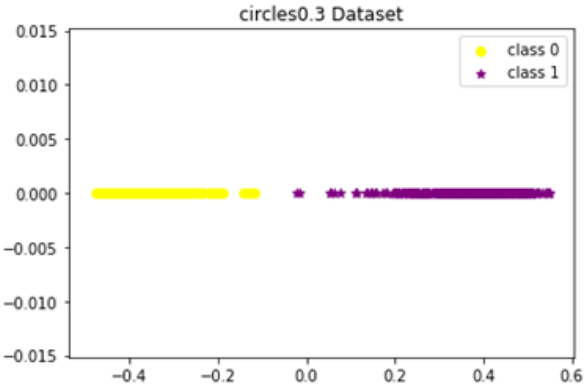

Fig. 10. Circles0.3 Dataset reduced to 1D by K-pca

and a classifier is trained on that data then the rest of the dataset is given as testing data to check the accuracy of the trained classifier. We use four classification algorithms including Naive Bayes, K Neareast Neighbor, Random Forest Classifier and Support Vector Machine on the dataset.

- **Naive Bayes Classifier:** The Naïve Bayes classifier [3] is based on the Bayes theorem. Which is given by the formula

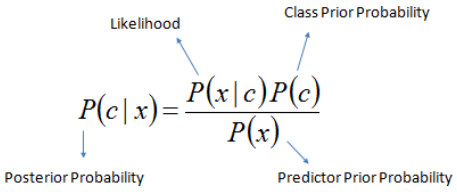

Fig. 11. Bayes Theorem

This classifier takes into consideration the probability of a datapoint (x) belonging to a particular class (c). This probability value is calculated for all the classes and the datapoint then becomes a member of the class that has the highest probability for that particular datapoint.

- **K Nearest Neighbor Classifier:** The K nearest neighbor classifier [2] (Fig: 12) works by calculating the distance of a new datapoint from the existing datapoints using a distance function. Then assigns a class to the new datapoint based on the majority class of the k nearest neighbors, it is desired that the value of k should be odd to avoid ties. In figure the new datapoint is assigned to the blue class.

- **Random Forest Classifier:** The Random forest classifier (Fig: 13) divides the dataset into random subsets and creates decision trees form each subset. It counts the result of each tree as votes to decide the final class of the datapoint.

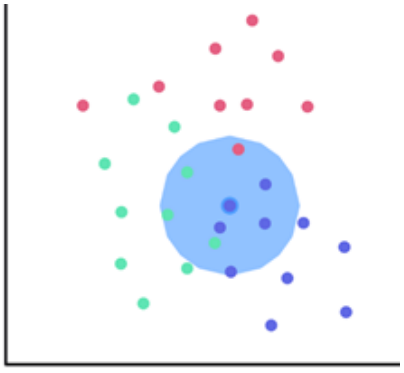

Fig. 12. K-Nearest Neighbor

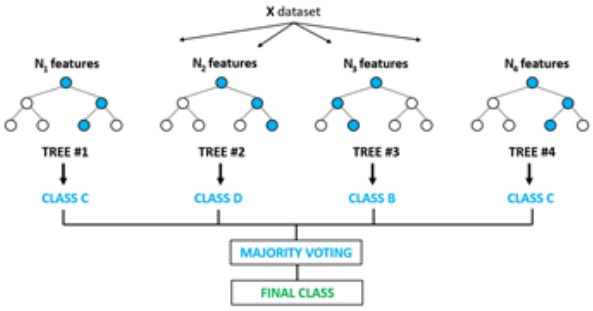

Fig. 13. Random Forest Classifier

- **Support Vector Machine:** A support vector machine [4] works by mapping the datapoints onto the N-Dimensions and creating a hyperplane that distinctly classifies the datapoints. The objective is to figure out a plane that maximizes the margin, i.e. the maximum distance between data points of both classes. This ensures that the future datapoints will be classified correctly. The points closest to the hyperplane are known as support vectors. (Fig: 14) shows how support vectors influence the orientation of the hyperplane. Additionally, selecting the optimal kernel function and tuning the parameters produces results that have high accuracy.The RBF kernel is the most powerful and is utilized when the dataset is complex and overlapping. The formula for the radial basis function is given below:
  The RBF Kernel uses the formula to map the data onto infinite dimensions and produces a hyperplane that

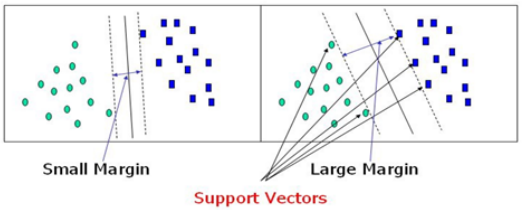

Fig. 14. Support Vector Machine

$$k(x,y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$$

Fig. 15. Radial Basis Function

separates the data.

b) Clustering: Clustering is the task of grouping the data points together such that the points in the same group are similar to each other. We have used two techniques for clustering, Expectation Maximization and Spectral Clustering to cluster the points in the Breast Invasive Carcinoma dataset.

- **Expectation Maximization:** This is a powerful algorithm as it has the ability to deal with missing data and unobserved features which makes it useful in many real-world applications. When my data is incomplete then Expectation-Maximization (EM) algorithm is a way to find maximum-likelihood estimates. This algorithm contains 2 steps, first is the ”E-step”. In this step, an initial guess is made for the model’s parameters and a probability distribution is created. Then, this model is fed with the new observed data. The next step is the ”M-step”. In this step the probability distribution is adjusted to include the new data. These steps are repeated until the distribution doesn’t change from the E-step to the M-step.

- **Spectral Clustering:** We have used the RBF kernel of spectral clustering, so number of neighbors parameter is ignored (in reference to sklearn documentation). In spectral clustering, the affinity, and not the absolute location (i.e. k-means), determines what points fall under which cluster. Spectral clustering is very useful in tackling problems where the data forms complicated shapes. In rbf kernel firstly a similarity graph is constructed in a higher dimension. Then, Adjacency matrix, Degree matrix and the Laplacian matrix are created. After this, eigenvectors of the matrix L are computed. Further, a k-means model is trained using the second smallest eigenvector to classify data. The function used while creating the affinity matrix when we use Gaussian or RBF kernel is given by

$$k(x,y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$$

Fig. 16. Radial Basis Function

distance(x,y) is the euclidean distance. This kernel in spectral clustering is useful when the clusters are not in the form of clouds in Euclidean space. In other words, when the clusters cannot be determined using the centroid. This algorithm is very useful when the clusters form a complex shape like say, nested circles.

V. Results Achieved

A. Classification

—

- **Naïve Bayes Classifier:** There are 3 distributions of Naive Bayes, namely, Gaussian, Multinomial, and Bernoulli. Multinomial Naive Bayes gives the best accuracy when there are a lot of features and the data of these features is in the form of frequencies. This means, that we must use Multinomial Naive Bayes in case of text classification. For all the below datasets, we have just 2 features. So, we cannot use Multinomial Naive Bayes. Bernoulli distribution is used when the data in features is in the binary form. So, for my below datasets, none of them have data in binary form. As a result, we do not use Bernoulli distribution on the below datasets. The Gaussian Probability function is given by

$$f_g(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-a)^2}{2\sigma^2}}$$

Fig. 17. Gaussian Probability Density Function

a is the mean of the values of the feature x, sigma is the standard deviation of the feature x. When the graph between the Gaussian probability function and x is plotted(figure below), it is seen that probability function reaches its maximum when x=mean value. This means that my Gaussian Naive Bayes algorithm will perform best when my data is around the mean value of that feature.

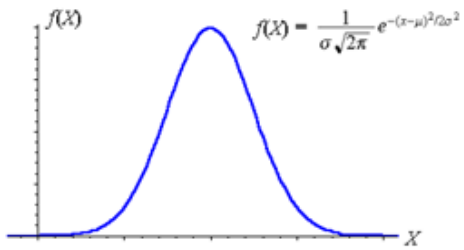

Fig. 18. Normal Distribution

So, due to the above reasons, I will be using Gaussian Naive Bayes on all the below datasets. As we are aware that the data present in Breast Invasive Carcinoma dataset contains more than twenty thousand features of 817 patients. But the values in each feature are not around the mean value of that feature. So, the when we use Naive Bayes classifier on this dataset for classification, the accuracy of the predictions is poor. In other words, we can say that Naive Bayes performs really well when the data is present around the mean value of that feature.

- **K-Nearest Neighbor:** The K nearest neighbor algorithm makes an assumption that the data points that are similar to each other always lie close. In this algorithm, we get a testing point that needs to be classified. The distance of this testing point from all the training point is calculated. These values are stored in a list. The list is sorted and the top K points are chosen. The majority of classes of all these K points is taken as the answer for the testing point. But, one major issue of KNN algorithm is the time complexity. As we know, the training time is negligible as no training is required by the algorithm. All the computations are done during the testing phase. So, this leads to bad user experience as no user wants to wait so long in order to see the classification result. Also, as our Breast Invasive Carcinoma contains many features, so it is really hard for the KNN algorithm to find the distance between any two points existing in different dimension. So, this is the reason why KNN performs poorly on our dataset.

- **Random Forest:** A decision tree is used for classification in which at each step a division takes place by using certain parameter. It consists of a Node, where a value is tested for splitting, edges, used to connect nodes with each other and leaf nodes that are the terminating nodes of a decision tree. We use classification trees for the purpose of classification. This tree is built using a technique called recursive partitioning. This is a process in which a node takes a decision of splitting into further branches based upon the feature value. This process is also divide and conquer because the data is spit into smaller parts in each step. When we use the decision tree as a classifier, we pick up a feature from the dataset showing the maximum information gain to represent the root node. We keep on picking the features from the dataset iteratively based upon the information gain till a pure node is created in the tree. A combination of such trees is called a random forest. The major advantage of this technique is that it is very fast in classification. But, overfitting can take place due to which the testing accuracy will be affected. Moreover, when the number of features are very large, decision trees can become very complex. They are difficult to understand and the results are inaccurate due to increased complexity. This affects the classification accuracy. Due to this reason, Random Forest classifier does not perform well on our Breast Invasive Carcinoma dataset.

- **Support Vector Machine:** In SVM, if we use linear kernel then no change in dimension takes place. So, no new features are created in this kernel. So, if we apply linear kernel we will simply get a 2-dimensional line as our decision boundary. As a result linear kernel will give less accuracy when the data is not linearly separable. According to the RBF kernelin (Fig: 19), where x is the current data point in n-dimension and x' is the some other data point.The square difference between x and x' is the Euclidian distance between the two data points and

sigma is the parameter that specifies the spread of the kernel. The RBF kernel represents the feature space in infinite dimension, so this is the reason why it gives such high accuracy when used for classification. Because of this reason, we will be using rbf kernel on the Breast Invasive Carcinoma dataset for classification. The figure below shows SVM with RBF Kernel classification on 8000 best features. The clusters have a reasonable amount of separation and the classifier achieves 97.8% accuracy on the dataset.

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right)$$

Fig. 19. Radial Basis Function

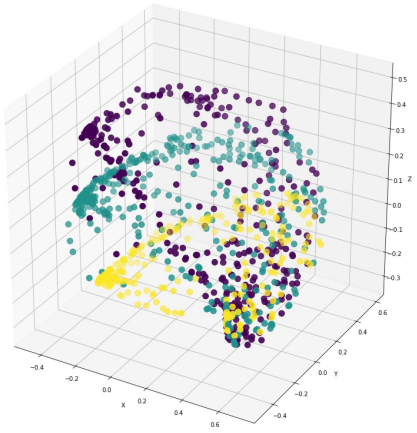

Fig. 20. Support Vector Machine Classification with RBF Kernel on IHC HER2

TABLE I
ACCURACY OF CLASSIFICATION ALGORITHMS

Algorithm	IHC HER2	ONCO-Tree Code
Naive Bayes	48.5%	53.5%
K-Nearest Neighbor	56.8 %	74.4%
Random Forest	82.9%	89.7%
SVM - Linear	87.3 %	83.8%
SVM - RBF	97.8 %	99.3%

B. Clustering

—

- **Expectation Maximization:** Expectation-Maximization (EM) is a powerful algorithm as it has the ability to deal with missing data and unobserved features which makes it useful in many real-world applications. When my data is incomplete then Expectation-Maximization (EM) algorithm is a way to find maximum-likelihood estimates. This algorithm contains 2 steps, first is the ”E-step”. In this step, an initial guess is made for the model’s parameters and a probability distribution is created. Then, this model is fed with the new observed data. The next step is the ”M-step”. In this step the probability distribution is adjusted to include the new data. These steps are repeated until the distribution doesn’t change from the E-step to the M-step. Although EM algorithm can provide good results when used on Breast Invasive Carcinoma dataset but there are some issues while clustering when this algorithm is used. As our Breast Invasive Carcinoma dataset is very large, so this algorithm shows slow convergence rate. Also, one major problem while using this algorithm is that it converges at local optima.

- **Spectral Clustering:** We have used the RBF kernel of spectral clustering, so number of neighbours parameter is ignored (in reference to sklearn documentation). In spectral clustering, the affinity, and not the absolute location (i.e. k-means), determines what points fall under which cluster. Spectral clustering is very useful in tackling problems where the data forms complicated shapes. In rbf kernel firstly a similarity graph is constructed in a higher dimension. Then, Adjacency matrix, Degree matrix and the Laplacian matrix are created. After this, eigenvectors of the matrix L are computed. Further, a k-means model is trained using the second smallest eigenvector to classify data. The function used while creating the affinity matrix when we use Gaussian or RBF kernel is given by numpy.exp(-gamma*distance(X,X)2) distance(X,X) is the euclidean distance. This kernel in spectral clustering is useful when the clusters are not in the form of clouds in Euclidean space. In other words, when the clusters cannot be determined using the centroid. This algorithm is very useful when the clusters form a complex shape like say, nested circles.

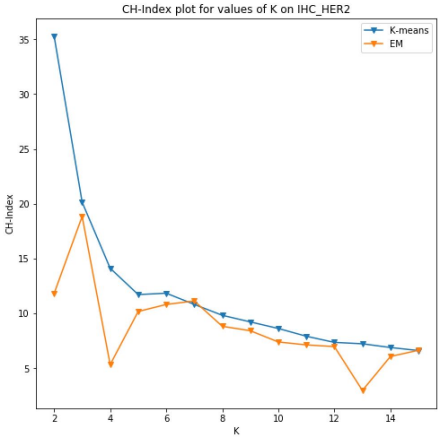

Fig. 21. CH-Index for EM and K-Means

K-means algorithm does not perform well on the dataset and no value of k can be determined using the elbow method on the graph in (Fig: 21). Using Elbow method on the EM graph determines that the value of k is equal to 3. This k-value is later applied in the clustering algorithm.

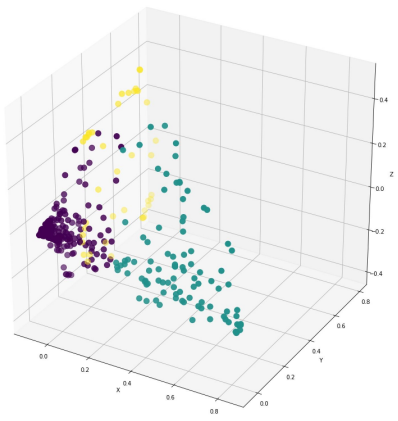

Fig. 22. Spectral Clustering with k = 3

VI. Ethical, Legal and Societal Aspects

A. *Ethical Aspects:*

Our project is designed as an aid to an oncology department of a hospital as it will help the doctor make an early diagnosis based on the gene data if the patient. That might indicate if the patient is HER2 positive or not. As a HER2 positive cancer is aggressive and the need for an emergency operation is determined by its value being positive or negative. This product is designed to provide an early warning for the oncologist as the cancer is a time sensitive disease and the prediction of a patient being HER2 positive might prove to be lifesaving. The time difference instant result of the classifier compared to the time taken for the actual test to be performed is large. But the classifier might also produce some level of inaccurate results. So, there is still a need for an actual IHC HER2 test to performed. The confidence in the classifier's ability to predict the HER2 result is high but not 100 percent. Over time this level may increase as the classifier is trained on new data. But this classifier will not replace a lab test.

B. *Legal Aspects:*

The owner of the data are the people that the data is about. we need to take necessary steps that ensure that the proper permissions are granted by the owners of the data. There is still a debate about the sharing monetary profits of the product with the people that have permitted the use of their data for the development of the product. if a potential life saving product is developed by people's data who are going through battling cancer in their own lives as well, we believe that they deserve a portion of the profit made by the product.
The accountability issue of a false positive or a false negative does not rest with the product as the prediction accuracy is high but not perfect. Hence, the doctor still responsible for making sure with the traditional tests that the patient is in actual condition the classifier predicted.

C. *Societal Aspects:*

This product is designed for all people and does not have a bias. Naturally its a type of cancer which more commonly effects women than men. Similar implementation can be done for prostate cancer which is a cancer that predominantly effects men. The product is designed for people with a terminal illness that can prove to be fatal. Helping save lives with machine learning will have positive impact on the society and further motivate people to use machine learning for the greater good and serve all people without any bias. Our product will not replace the job of the doctor or the person that is responsible for the IHC Her2 test. this product will only help the doctor make a prediction potentially helping the doctor save a life.

VII. Conclusion

Using the dataset that initially had 22247 features for 817 patients, We have selected 8000 most important features from the dataset while making sure that there is no loss of important information while selecting the features. We have achieved an accuracy of 97% on the predicting the result of the Immunohistochemistry HER2 test result that determines the aggressiveness of Breast Invasive Carcinoma.

We have also discussed the ethical and societal implications of this machine learning research project and how it will enable the doctors save lives just with the genetic information of the patient diagnosed with the cancer.

References

[1] Furey, T. S., Nello, Duffy, Nigel, Bednarski, D. W., Schummer, Haussler, and David, "Support vector machine classification and validation of cancer tissue samples using microarray expression data," OUP Academic, 01-Oct-2000.

[2] Srivastava, T., & Srivastava, T. (2019, September 3). Introduction to KNN, K-Nearest Neighbors : Simplified. Retrieved from https://www.analyticsvidhya.com/introduction-k-neighbours-algorithm-clustering.

[3] www.statsoft.com Retrieved from http://www.statsoft.com/textbook/naive-bayes-classifier.

[4] What is a Support Vector Machine (n.d.). Retrieved from https://www.kdnuggets.com/2017/02/yhat-support-vector-machine.html.