Using Proteogenomics Data to Determine Predictors of Breast Cancer and Its Determinant Conditions

Leonardo Vieira, Cheronika Donald, Ying Zhang, Yemi Osho, and Fang Chen
Jackson State University

In the healthy body, natural systems control the creation, growth, and death of cells. When these systems do not work properly cell growth outweighs cell death and cancer occurs. That excess growth can form a tumor.

Breast are made up of glands called lobules that can make milk and have thin tubes called ducts that carry milk to the nipple. Breast tissue also contains fat and connective tissue, lymph nodes and blood vessels. Some breast cancers can begin in the cells of the ducts, cells of the lobules or in other tissues in the breast. This research will explore prognosis and diagnosis of breast cancer based on clinical data. Factors such as age, tumor stage, metastasis and response to chemotherapy will also be explored. Estrogen receptors (ER), Progesterone receptors (PR), and Human epidermal growth factor receptor-2 (HER2), are part of the predictive factors that will be used: they provide information on the likelihood of response to therapy administered to the subjects.

At the end of this research, a machine learning tool will be constructed to predict if subjects have a likelihood of having cancer, which stage of cancer, and the likelihood of the subject positively responding to treatment.

1. Introduction

Though the data chosen to complete this research has been tested with K-means, we will test other machine learning methods to see if a better model can be implemented and give better results.

We propose using several different machine learning prediction model that we can: a) clustering tumor apply to chemotherapy treatment on protein data, b) predict age on both protein data and clinical data, c) predict tumor stage, Node stage and Metastasis progress by using protein data and clinical data, and d) predict AJJC stage by using protein data and clinical data.

2. Data

The dataset contains published iTRAQ proteome profiling of 77 breast cancer samples generated by the Clinical Proteomic Tumor Analysis Consortium (NCI/NIH). It contains expression values of samples for approximately 12,000 proteins, with missing values present when a given protein could not be quantified.

There are three contents of this dataset:

(1) 77_cancer_proteomes_CPTAC_itraq.csv; this file contains: RefSeq_accession_number, which is RefSeq protein ID (each protein has a unique ID in a RefSeq database); gene_symbol which indicate a symbol unique to each gene because every protein is encoded by some gene; gene name, which is a full name of that gene for the remaining columns: log2 iTRAQ ratios for each sample (protein expression data, most important), three last columns are from healthy individuals. The positive values of the expression data denote a simulation action in the person and the negative values of the expression data denote an inhibition action.

(2) clinical_data_breast_cancer.csv; included is clinical data on two male and 103 female subjects. This dataset has 30 features which breakdown as follows: first column, "Complete TCGA ID", is used to match the sample IDs in the main cancer proteomes file. All other columns have self-explanatory names, contain data about the cancer classification of a given sample using different methods. The main features referenced in this dataset are Age at Initial Pathologic Diagnosis, ER Status, PR Status, HER2 Final Status, Tumor, Node, Metastasis, AJCC Stage, Converted Stage, Survival Data Form, OS event, and PAM50 mRNA (these will be discus later in the study).

(3) PAM50_proteins.csv; this is a list of genes and proteins used by the PAM50 classification system. The column RefSeqProteinID contains the protein IDs that can be matched with the IDs in the main protein expression data set.

3. Methodology

Methodology contains data preprocessing, find best model to train the data, divide data into train fold and test fold. data visualization

In the data preprocessing we sought to: understand the data, analyze the data, which include correlation, analyze the structure, and percentage of missing data of the datasets. Next, we reconfigured the data frame by filling in the missing data, mapping, reducing, and using principle component analysis to normalize the data.

1. Understanding the data: in the previous study researchers demonstrate that proteogenomic analysis of breast cancer elucidates functional consequences of somatic mutations, narrows candidate nominations for driver genes within large deletions and amplified regions, and identifies therapeutic targets. We followed the previous paper using the new proteomics data. Proteomics is the study of the protein which can help us to target the cancer candidate more

accurate. In our study, we want to combine the proteomics data with clinical data for the breast cancer patient and with PAM50, which is the list of the cancer biomarker to find the correlation and use them apply in the predictive model to solve regression, classification and clustering problem.

2. We mapped some of the features from the clinical data to be used and drop some of the features which have the same function as others such as Tumor and Tumor—T1 Coded, Node and Node-Coded, Metastasis and Metastasis-Coded, etc. These features have similar function for example, in Tumor and Tumor—T1 Coded, the Tumor denote the Tumor stage 1 to 4, and the Tumor—T1 Coded only denote patients has Tumor stage 1 or not. Age at Initial Pathologic Diagnosis will be used as one of our targets to predict the patients age according to the protein data and clinical data. ER status denotes the person is ER positive or negative, the patients who are ER positive tend to have better chemotherapy treatment with anti-hormone drug which can block estrogen from attaching to the breast cancer cells. ER negative person will have better treatment with regular chemotherapy treatment according to previous clinical research. PR Status also denotes your hormone level but for Progesterone receptor can denote a different chemotherapy treatment. HER2 (human epidermal growth factor receptor 2) is one gene that can play a role in the development of breast cancer. HER2 final status tells whether or not HER2 is playing a role in the cancer. The HER2 gene is also called the ERBB2 (Erb-B2 receptor tyrosine kinase 2) gene, so it may be referred to by that name in some studies. HER2-positive breast cancers tend to grow faster and is more likely to spread and come back compared to HER2-negative breast cancers. But there are medicines specifically for HER2-positive breast cancers, which also show the difference in treatment of various chemotherapies. Tumor, Node and Metastasis denote an important staging called TNM system. In the system, Tumor denotes the size of your tumor. The larger the number the larger the tumor size. Node denotes the number and location of lymph nodes that contain cancer. The higher the number after the N, the more lymph nodes contained in the cancer site. Metastasis denotes two stages which are: M0; Cancer has not spread to other parts of the body, M1; Cancer has spread to other parts of the body. TNM system will be used as one of our targets to classify the patient and used to predict patient's staging. AJCC Stage is a classification system developed to utilize, in part, the TNM scoring system: Tumor size, Lymph Nodes affected, Metastases. This will also be used as a target feature. Converted Stage denotes if and how the subject's cancer state improved. Survival Data Form denotes subject enrolled in follow-up or had follow-up. Remaining features in the *clinical data breast cancer* dataset are values generated from mass spectrometry (analytical technique that measures the masses within a sample).

3. The *77* cancer proteomes CPTAC itraq dataset consists of reference sequence accession numbers, the gene symbol, and the gene name for each section of protein studied. The remaining 83 features consist of mass spectrometry values (80 features represent subjects from the clinical data breast cancer dataset). These 80 values tell if the subject has that section of protein present and the amount of that protein in the tissue sample collected.

4. Missing data: Python pandas was used as our tool to analyze the data and find percentage of missing data in the proteomics data (cannot be simply labeled zero because the spectral analysis indicated equal environment when the number is zero). Missing data indicated the mass spectrometry did not have the ability to scan the protein's reaction. Therefore, we fill in the missing data using medium number.

5. Map and Reduce: Based on the proteomics data and PAM50, which is the protein name of cancer protein candidate, we map the proteomic data by filtering with PAM50 data. The result was that the data contains patients and their cancer protein candidate expression data. In this process, there were some PAM50 proteins missing. Then we reshape the data using patients as the instances and using protein expressions as features. Therefore, we can clearly see the patients as each instance using cancer protein candidate expression data as attribute to do the next work. Then we merged the data with clinical data we have for the patients. The goal helps us to filter out some of the patients without clinical data, which keeps the consistency.

6. Spilt the data: For comparison purposes, we split the data again into new protein expression data and new clinical data. The purpose is to clarify the data in the same format. There is no literature that showed protein expression data being simply combined with the clinical data is a good way to use as a predict model.

7. Principle Component Analysis (PCA): PCA was used to reconfigure the protein expression data. The protein expression data had an enormous number of attributes. PCA will help in decorrelation and dimensioning the data. It will also help to select the column which has the most information. Although, the biggest component number is very small, it still helps to arrange the protein data.

Research Goal:

1. Using patient's cancer candidate proteinomics data to cluster the protein to see whether we can use the clustering as a sign of tumor stage, or we can cluster the portion that can have different protein treatment.

2. Using age, Tumor, Node, Metastasis, AJCC stage as targets to predict.

Model and result:

The criteria to choose our Methods and Models was made by talking to our domain experts Dr. Sravanthi Joginipelli and Dr. Venkata Melapu, who were able to certify that we will be able to answer our objective questions using those techniques.

1. Clustering: base on the limited knowledge we have to handle protein expression data, we do not have any target data to predict. So, we use unsupervised learning as the technique to analysis the protein expression data. We use K-mean algorithm to cluster the protein data. The K-means is a clustering technique that finds *n* observations into *k* clusters where each observation belongs

to a different cluster based on their nearest mean. We use Tumor stage in the patient clinical data to test whether it has correlation with our result. Test showed we can clustering the protein data when K = 3, but there is low correlation with the tumor stage.

2. Regression: Second task is to predict the age base on both clinical data and protein expression data. Since age is numerical data, we used it as target data. We extract the target data from clinical data and normalized the remaining clinical data. We then used multiple linear regression, Support vector regression, Kernel support vector regression, decision tree regression and Random forest regression to predict the result.

3. Classification: Third task is to predict the Tumor stage, Node stage and progress of metastasis, AJCC stage and OS event based on both clinical data and protein expression data. Tumor stage has four different stages, Node has three different stage and Metastasis has 0 and 1, therefore, we extract the target data from clinical data each time and normalize the rest of the clinical data. Then applied Logistic regression classification, k-nearest neighbors, support vector machine, kernel support vector machine, Naive Bayes classification, decision tree classification and random forest classification. Logistic regression and SVM only can predict two targets, which is 0 and 1. So we applied these only to progress of metastasis. With the other techniques, we applied all of our targets. We selected radial basis function kernel in kernel SVM to predict all targets. Entropy was used as a key feature in determining the decision tree to predict targets. Also in the random forest classification, we used n_estimator as 10 and criterion = ‘entropy' to predict the target. Result will show our outcome on the graph compared to the test data.

4. ANN: Finally we use ANN(artificial neural network) to predict all of our targets. We use tensorflow as our work backend, Keras as the interface, hidden layer using the number of the targets and the number of the input value divided by two, (for example, we have 4 targets, and we have 43 input value in our protein data, so our hidden layer is (4+43)/2 = 24). We chose rectifier function in the input layer and multiple sigmoid function as our output layer, which activation function is called softmax. We also calculate the loss function and accuracy by using the sparse categorical crossentropy function. Results will show our outcome on the graph compared to the test data.

Discussion

Base on the graph our error is very high compare to test target. We feel there are three ways we can make the error lower in the future.

1. Handling the missing data: We used medium as a method to fill in the missing data. It was suggested we should use shannon diversity index to normalize and fill in the missing data, but as we were processing it, we found that shannon diversity index only work with positive numbers. Our protein expression data included negative numbers as action of inhibition. So, we need to read more literature review on the how to handle the missing data from protein expression data.
2. Normalize the data: As we mention above, we also should use shannon diversity index as the method to normalize the data. PCA can give us some clear information about how the data

performs and provide a very nice way to handle the data, but our data after reshaping using PCA, the first column only in put 0.0016% of the whole data information. Therefore, PCA actually is not a good way to normalize and reshape our protein data. We also need further research on how to normalize the protein expression data.

3. The prediction model we use: We used transitional machine learning model as our learning processing. All of the models used overfit the data when we used the model to learn not only the protein expression data but also the clinical data. In our traditional method, we all apply weight into the learning, but in the gene and protein data learning, some of the paper should the in-correction of the way of doing so. In the future we should study un-weighted learning and apply in our protein expression data and clinical data.

Principal Results:

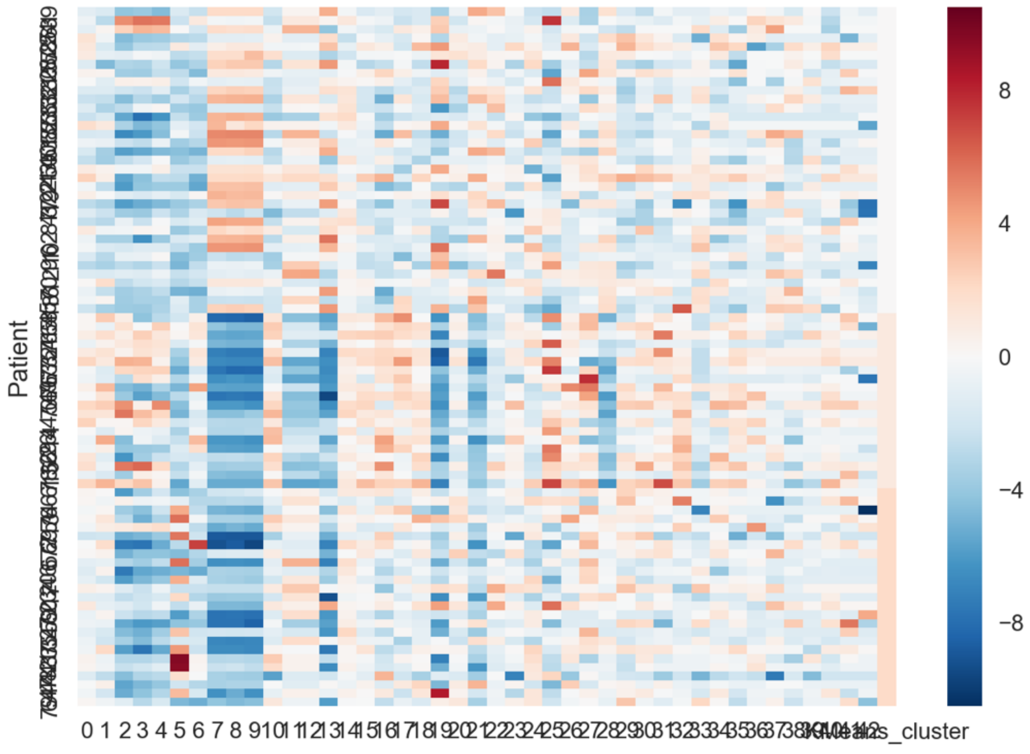

Fig 1. Correlation between patient and proteins.

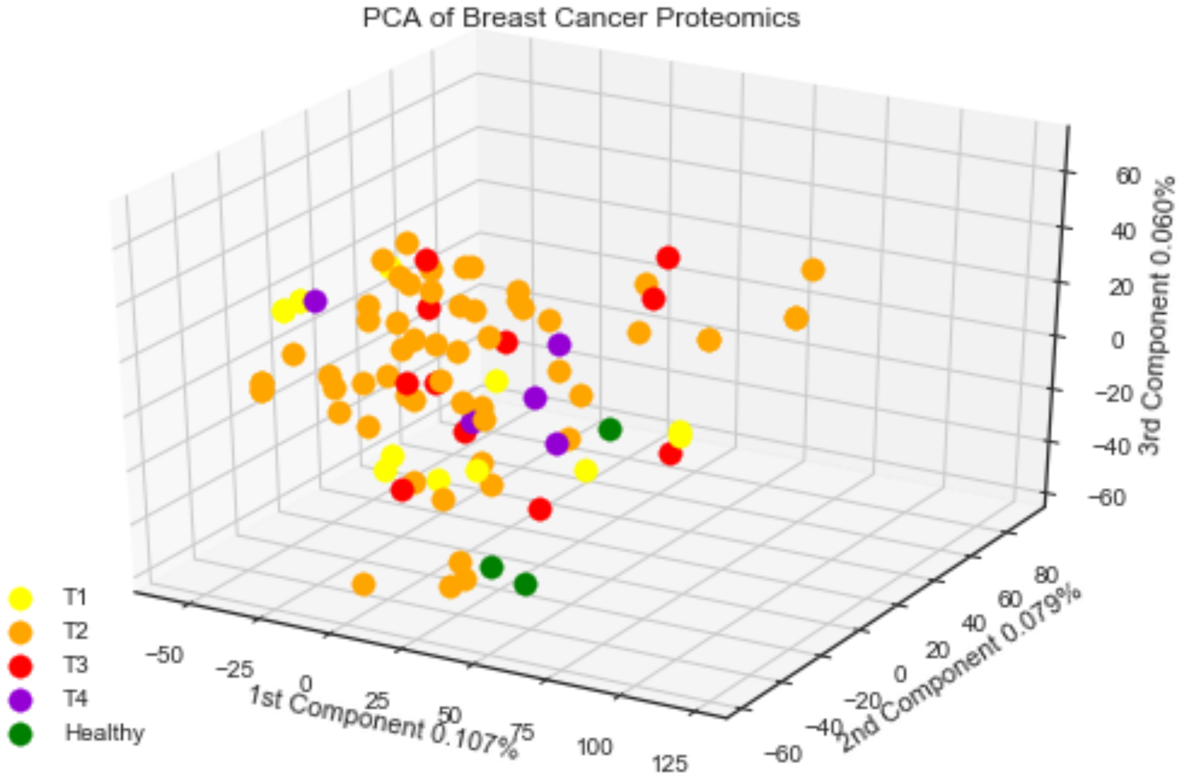

Fig 2. Principal component analysis of the 3 proteins with more information.

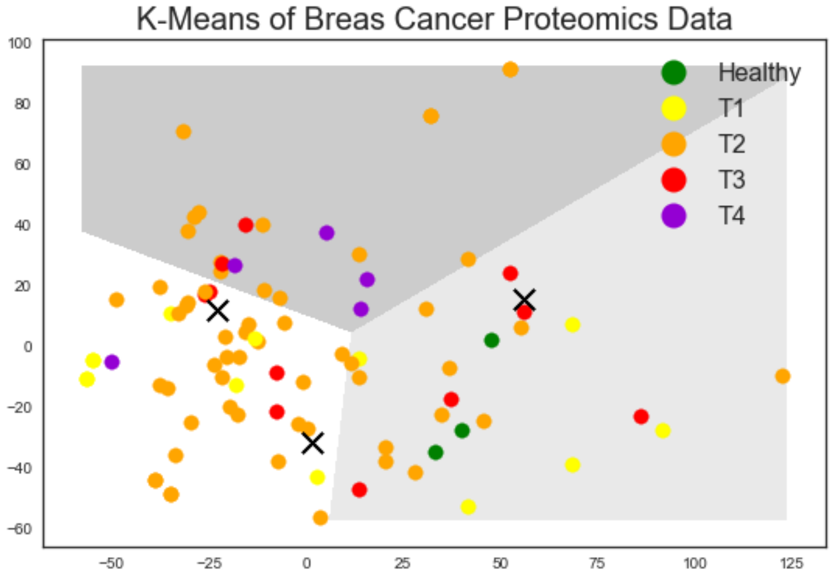

Fig 3. K-Mean cluster using the proteomics data.

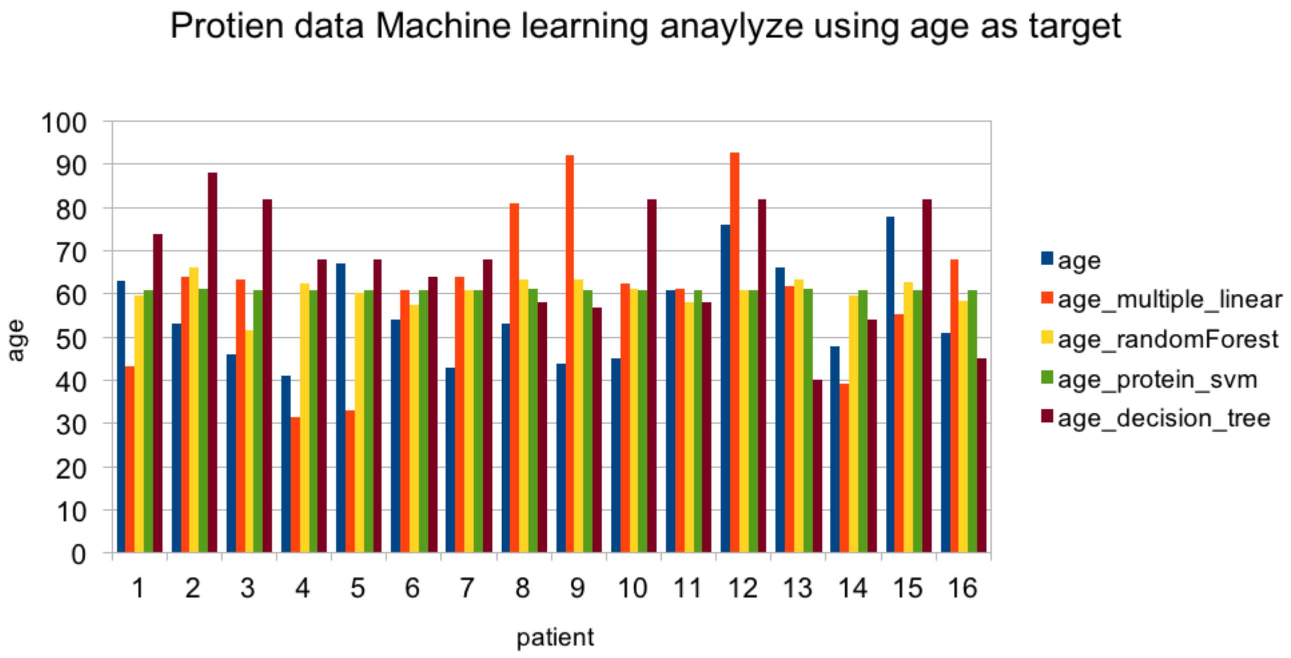

Fig 4. Age based on proteomics data.

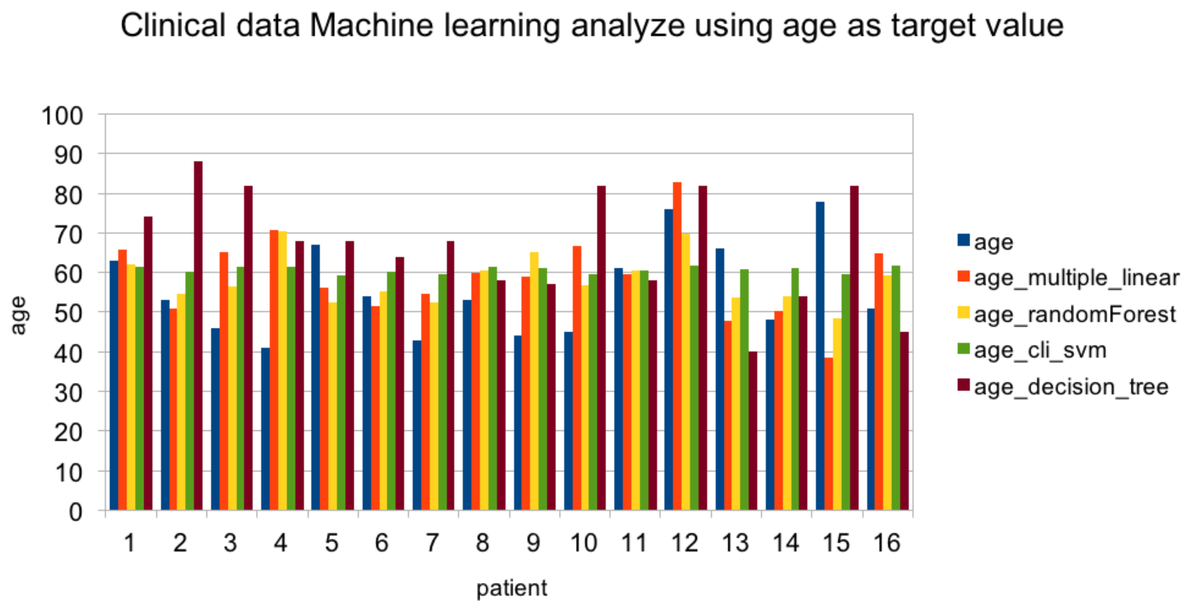

Fig 5. Age based on clinical data

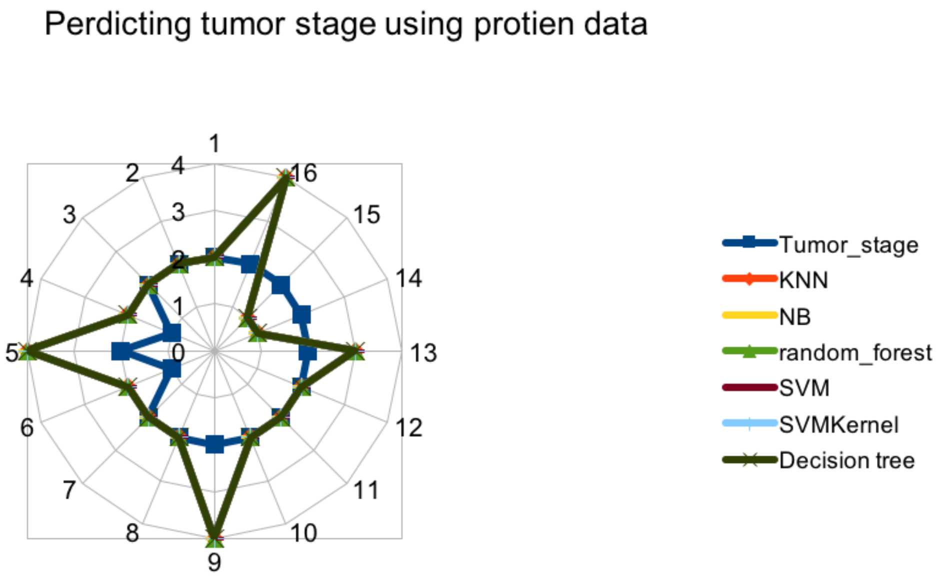

Fig 6. Tumor prediction based on proteomics data.

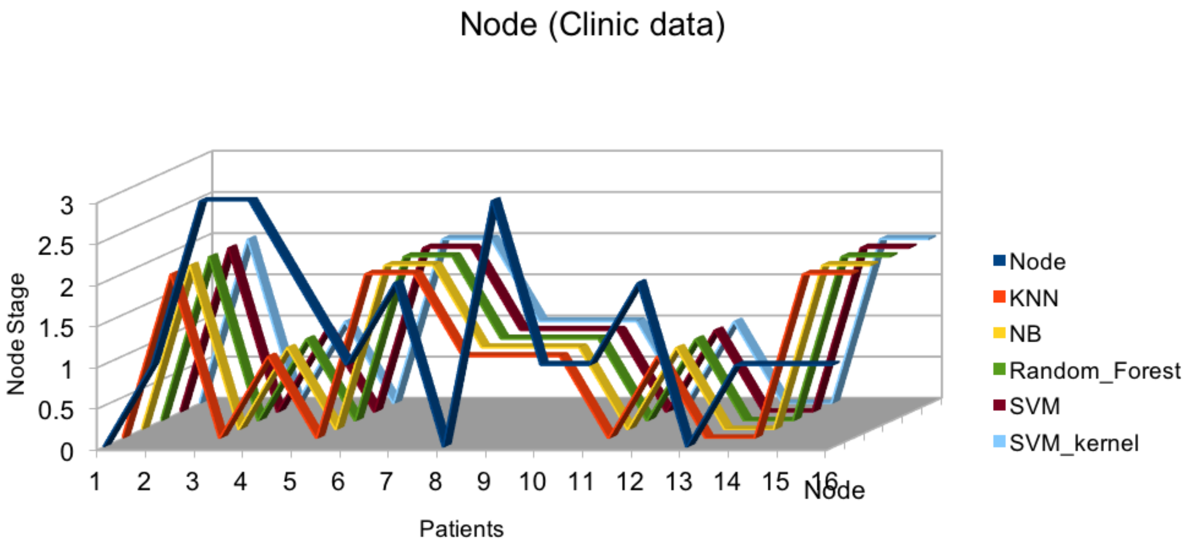

Fig 7. Node stage based on clinic data.

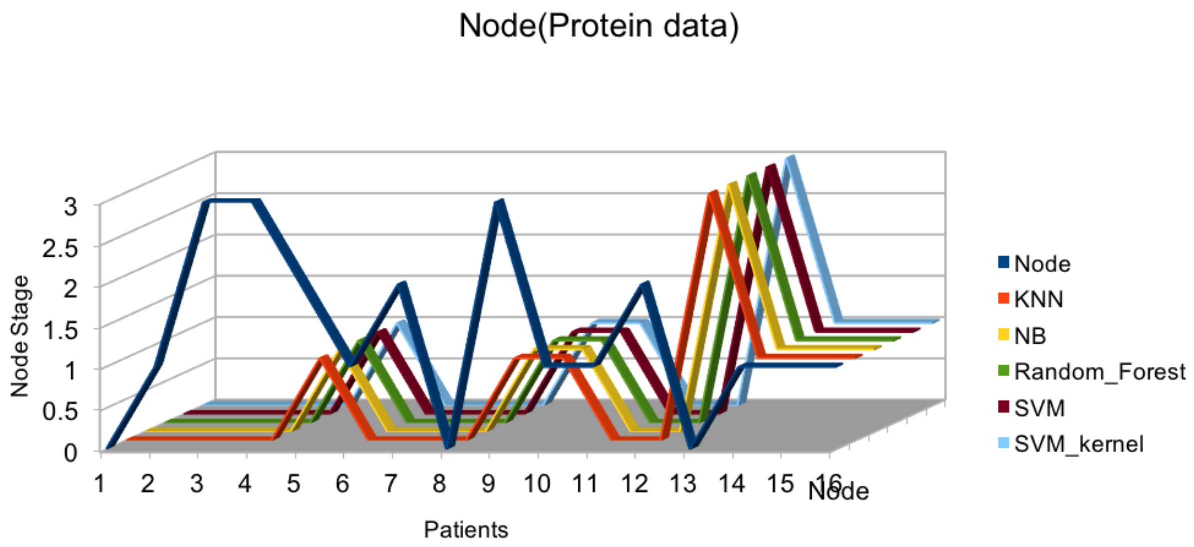

Fig 8. Node stage on proteomic data.

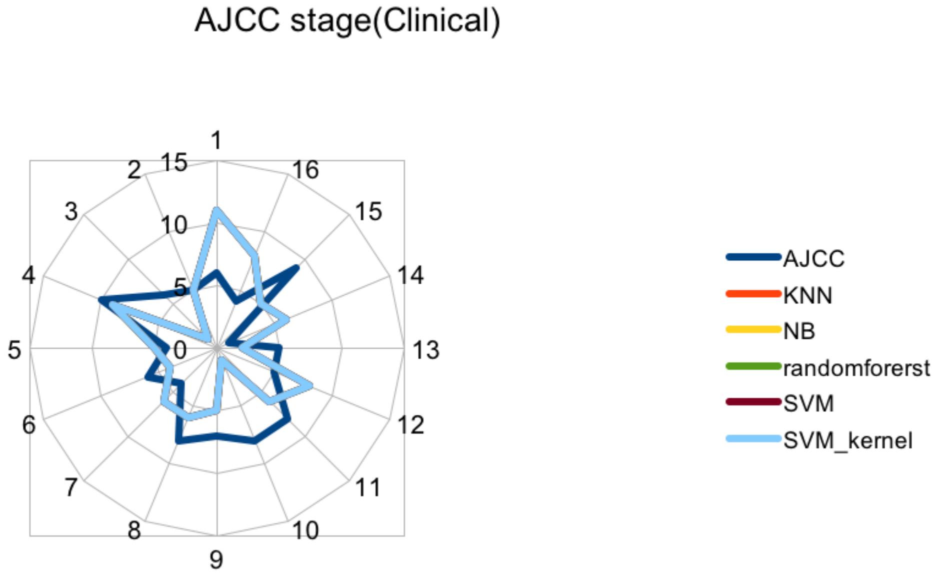

Fig 9. AJCC stage based on clinical data.

10. Bibliography

[1] Hearst M.A., Schölkopf B., Dumais S., Osuna E., and Platt J. 1998. Trends and controversies-support vector machines. IEEE Intelligent Systems 13: 18-28.

[2] Lugendra Dongre, Gend Lal Prajapati, "The Role of Apriori Algorithm for Finding the Association Rules in Data Mining", *ICICT*, pp. 657-660, 2014.

[3] Alan O. Sykes. An Introduction to Regression Analysis. www.law.uchicago.edu/files/files/20.Sykes_.Regression.pdf

[4] *Susan G. Komen*. (2016, 03 22). Retrieved from Facts and Statistics: http://ww5.komen.org/BreastCancer/FactsandStatistics.html

[5] Halls, D. (2017, February 21). *Moose and Doc Breast Cancer*. Retrieved from Moose and Doc Breast Cancer: http://breast-cancer.ca/6c-er-pr-her2/

[6] Health, N. C. (n.d.). *Cancer*. Retrieved from Cancer Types: https://www.cancer.gov/types/breast