CLASSIFICATION AND PREDICTION OF SURVIVAL STATUS AND CANCER PROGESSION OF BREAST CANCER PATIENTS

Zarreen Naowal Reza[1], Rajasi Upadhyay[2], Sumit Khairnar[3]
Department of Computer Science
University of Windsor
Windsor, ON N9B 3P4

***Abstract*—Breast cancer is a major health concern for women being a life-risking disease affecting the breast tissue. It often grows very slowly, often causing no significant symptoms until it reaches the advance stage. Although there are some symptoms of this disease like change in breast shape, red scaly patch of skin but, these are not easy to find at an early stage. Therefore, prognosis and diagnosis of this disease is a big challenge at an early stage. Biomarkers are really useful in overcoming this challenge as it can detect the presence and progression of cancer in the body. In this study, the survival status of breast cancer patients is classified in classes like *living* and *deceased* based on different combination of surgery and therapies they have received using gene expressions. Classification task is performed with Naïve Bayes, Random Forest and Support Vector Machine (SVM) classifiers. In addition, NPI score of patients is predicted based on different combination of surgery and therapies as well as five breast cancer subtypes. Regression task is performed with Random Forest Regressor, Support Vector Regressor (SVR) and Kernel Ridge Regressor (KRR). Different feature selection models are also used including mRMR and Regularized Linear Model. The performance of the models is evaluated based on ROC analysis obtained from 10-fold cross-validation and model selection procedure. According to final observation, the model with ElasticNet Regularized Linear Mode combined with Recursive Feature Elimination (RFE) provides the highest accuracy with the lowest number of features when classified with RBF SVM. For regression task, KRR predicts the NPI score most accurately with the lowest number of predictors.**

Keywords—Random Forest, SVM, Lasso, ElasticNet, RFE, PCA, LDA, gene expressions, Kernel Ridge, SVR, RFR

I. INTRODUCTION

The prospect of developing breast cancer is a source of anxiety for many women. Breast cancer remains the most common invasive cancer among women (aside from nonmelanoma skin cancers), accounting in 2011 for an estimated 230,480 new cases among women in the United States and another 2,140 new cases among men (ACS, 2011). After lung cancer, it is the second most common cause of mortality from cancer for women, with about 39,520 deaths expected in the United States in 2011. Another 450 breast cancer deaths are expected among men in 2011 (ACS, 2011). Since the mid-1970s, when the National Cancer Institute (NCI) began compiling continuous cancer statistics, the annual incidence of invasive breast cancer rose from 105 cases per 100,000 women to 142 per 100,000 women in 1999 (NCI, 2011). Since then, however, the incidence has declined. In 2008, the incidence of breast cancer was 129 cases per 100,000 women. Further reduction of the incidence of breast cancer is a high priority, but finding ways to achieve this is a challenge. One way to solve it is to use to data generated with the previous cancer patients. There is a large amount of data generated in curing the cancer patient. This data can be useful to make predictions for patients in future. Data collected on the basis of clinical observations serve as a raw data. The raw data is not at all useful for the future works but, we need to make it useful. Data analytics can be applied to this raw data such as machine learning techniques, data mining tools and also some patterns can also be drawn from the previous procedures to help and cure the patient. Different techniques have been used to aid clinicians in the estimation of the prognosis of different diseases. Standard statistical tools, like the Cox regression model or logistic regression are normally used. However, more sophisticated models, based on machine learning methods, have been applied in recent years. In many cases, they perform better than standard statistical tools. The advantage of using machine learning models is that they are more flexible (can be built on our own) than the standard statistical models and can capture more interactions between the data and result in better predictions.

The rest of this paper is arranged as follows. Section II discusses theoretical study. Section III explains the entire methodology. In Section IV result and experiments are discussed step by step. Section V gives some future research directions. Section VII concludes this paper with a brief summary.

II. THEORETICAL STUDY

Fig 1. demonstrates the axillary lymph nodes or armpit lymph nodes (20 to 49 in number) drain lymph vessels from the lateral quadrants of the breast, the superficial lymph vessels from thin walls of the chest and the abdomen above the level of the navel, and the vessels from the upper limb. They are divided in several groups according to their location in the armpit. These lymph nodes are clinically significant in breast cancer, and metastases from the breast to the axillary lymph nodes are considered in the staging of the disease. Lymph node status is the number of positive auxiliary lymph nodes observed at time of surgery. As shown in Fig 1, the lymph nodes in the armpit (the axillary lymph nodes) are the first-place breast cancer is likely to spread.

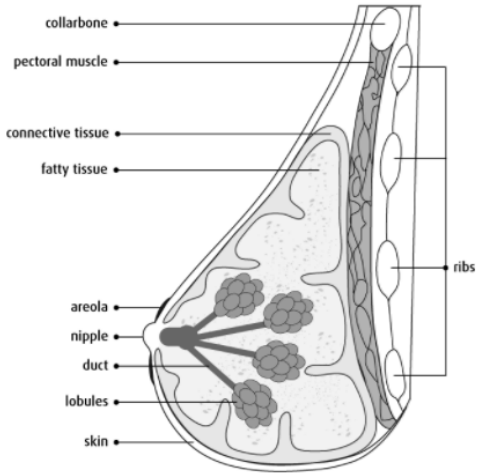

Fig 1: Breast Tissue [1]

Status of the axillary lymph nodes is highly related to the prognosis. Lymph node-negative means the lymph node does not have a cancer and lymph node-positive means the lymph node which has the cancer. Tumor Size is the diameter of the excised tumor in centimeters. Tumor Size is divided into four classes:

1. T-1: 0 – 2 Centimeters
2. T-2: 2 – 5 Centimeters
3. T-3: More than 5 Centimeters
4. T-4: Tumor of any size

Machine learning is one of the most important field of computer science which gives computers the ability to learn without being explicitly programmed. This involves the development of algorithms that learn how to make predictions based on data, has a number of emerging applications in the field of bioinformatics. Bioinformatics deals with computational and mathematical approaches for understanding and processing biological data. Prior to the emergence of machine learning algorithms, bioinformatics algorithms had to be explicitly programmed by hand for problems such as protein structure prediction which was extremely difficult. Later, machine learning techniques enable the algorithm to make use of automatic feature learning which means that based on the dataset alone, the algorithm can learn how to combine multiple features of the input data into a more abstract set of features from which to conduct further learning. This approach of learning patterns in the input data allows such systems to make quite complex predictions when trained on large datasets. In the breast cancer surgeries can be examined to see if it can cure the breast cancer or not. Since the decision involves whether the patient will be alive or deceased, it is required to find better classification and predication techniques to ensure that it gives accurate survival status.

For this research, Breast Cancer (BC) dataset containing 1980 samples or patients' reports are accumulated including 20 clinical variables (Biomarkers) and 24369 gene expressions. The dataset also consists of other features like Vital-status, NPI, Breast-surgery, Claudin-subtype, Chemo-therapy, Hormone-therapy, Radio-therapy, Histological-subtype. The vital status shows the condition of the patient as in Living, Died of Disease, or Died of Other Causes. The Nottingham prognostic index (NPI) is used to determine prognosis following surgery for breast cancer. Its value is calculated using three pathological criteria:

1. The size of the lesion
2. The number of involved lymph nodes
3. The grade of the tumour

In breast cancer, gene expression analyses have defined five tumour subtypes (luminal A, luminal B, HER2-enriched, basal-like and claudin-low), each of which has unique biologic and prognostic features, which is indicated in claudin-subtype for each patient. Chemotherapy uses anticancer (or cytotoxic) drugs to destroy cancer cells. Many women with breast cancer undergo treatments like Chemotherapy, Radiotherapy, Hormone Therapy and Breast Surgery. About the histological subtype field, it can be indicated that breast cancer is a genetically and clinically heterogeneous disease. In order to organize this heterogeneity and standardize the language, breast cancer classification systems have been developed. Classification of the dataset has evolved over time into a tool that is used to aid in treatment and prognosis. At first it was performed without using various models but now a days different models are available to perform the classification [6]. In this study, the Naïve Bayes, Random Forest and Support Vector Machine classifiers are used to classify patient's survival status and Random Forest Regressor, Support Vector Regressor and Kernel Ridge Regressor are used to predict NPI score of patients based on different treatment they have received. The experimental settings and simulation of dataset have been carried out using Python.

There are many types of breast cancer that can be identified based on the gene expression characteristics. It is very important that patient receives best therapy related to breast cancer where appropriate diagnosis of the specific subtypes can be done. In this study, patients receiving four categories of treatment have been taken into consideration.

A. Breast surgery

Mastectomy is the surgical removal of the breast. Types of mastectomy include simple, modified radical, and radical. The removal of some or all of the lymph nodes may or may not occur during the mastectomy. In most case reconstruction is possible almost immediately after the surgery. Mastectomy is usually the treatment for women with Stage 0, Stage I, Stage II, or Stage III level of breast cancers.

B. Chemo Therapy

Chemotherapy is an anticancer drug therapy and is normally given intravenously (through the vein) or orally in pill or liquid form. Chemotherapy may be used on its own or with a lumpectomy or mastectomy. Chemotherapy is a systemic form of treatment; it flows through the bloodstream, affecting the entire body. Its purpose is to obstruct the DNA synthesis of cancer cells. The appropriate combination of drugs used for

chemotherapy is based on the patient's cancer type and individual medical profile.

C. Hormone Therapy

Estrogen is a hormone produced by the varies, that gives rise to many breast cancers. Women whose breast cancers test is positive for estrogen is given hormone therapy treatment to lower estrogen levels. Tamoxifen and fareston are the 2 types of drugs that prevent estrogen from binding to breast cancer cells. Treatment of ER+ breast cancer with Tamoxifen for 5 years has been shown to reduce the rate of recurrence by 39%.

D. Radio Therapy

Radiation therapy is an adjuvant treatment for most women who have undergone lumpectomy and for some women who have mastectomy surgery. In these cases, the purpose of radiation is to reduce the chance that the cancer will recur locally (within the breast or axilla). Radiation therapy involves using high-energy X-rays or gamma rays that target a tumor or post-surgery tumor site.

III. METHODOLOGY

The entire study is divided into several stages, namely Feature Selection, Dimensionality Reduction, Classification, Regression, Cross-validation and Model Selection. For each stage, different relevant methods and algorithms are applied, individual performance is evaluated and finally comparison between the results are demonstrated.

A. Feature Selection

In Breast Cancer dataset, there are 24,369 predictors. Very high dimensional feature set are most likely to contain redundant information intervening with the determination of the true relationship between the features and the response variable [4]. In addition, working on very high dimensional dataset requires more computational time and cost. Therefore, picking the smallest subset of the most relevant features are of supreme importance. In this study, Minimum Redundancy Maximum Relevance (mRMR), Regularized Linear Model with Lasso (L1) and ElasticNet (L1 and L2) penalty and Recursive Feature Elimination (RFE) are used for feature selection purpose. Initially, three feature selection models are implemented. In the primary model, mRMR method is implemented on the entire feature set in order to extract the most relevant 50 features based on their individual association with the response variable. One disadvantage of this method is that neither it takes into account the correlation between features nor it does evaluate features by taking in groups. However, in Breast Cancer dataset some of the genes are highly correlated with each other and also correlated with the response variable as a subgroup. As a result, genes obtained from mRMR provide unsatisfactory outcome with average accuracy rate of 56%. Thus, this model is discarded and rest of the analysis has been carried on with the latter two models – Model A and Model B.

Model A proposes a feature selection technique which is a combination of two Regularized Linear Model regression methods called Lasso and ElasticNet followed by RFE. In this proposed method, top 50 features obtained from both Lasso and ElasticNet are combined as a feature subspace. After that, RFE is used on that subspace to rank the features starting from most to least important for predicting the response variable. Finally, top 15 to top 25 features are picked from the ranked order as the final feature subset.

On the other hand, Model B proposes a feature selection technique with only ElasticNet followed by RFE. In this proposed method, top 50 features obtained from ElasticNet are considered as a feature subspace. RFE is used on that subspace to rank the features starting from most to least important for predicting the response variable. Finally, top 15 to top 25 features are picked from the ranked order as the final feature subset.

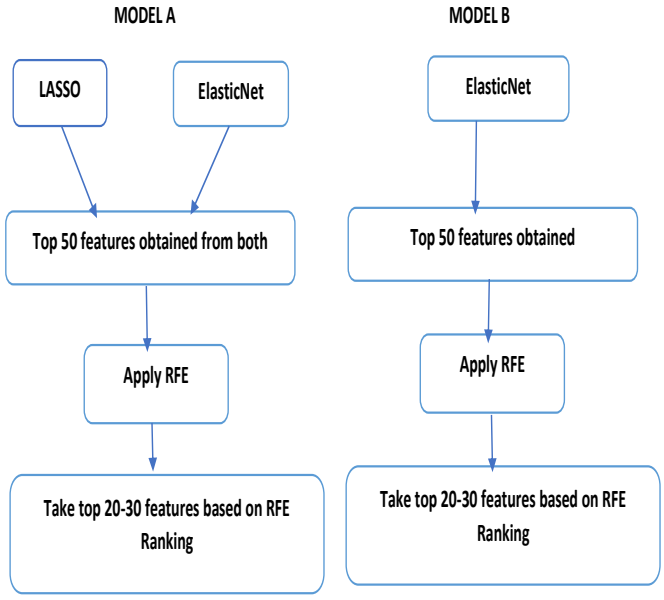

Fig 2: Flowchart of Model A and Model B

B. Dimensionality Reduction

In this paper, Principal Component Analysis (PCA) followed by Linear Discriminant Analysis (LDA) is applied as techniques of dimensionality reduction. PCA ignores class labels and its goal is to find the directions or principal components that maximize the variance in a dataset. In contrast to PCA, LDA computes the directions or linear discriminants that represent the axes that maximize the separation between multiple classes. After applying feature selection, 15-25 features are obtained. Then, PCA followed by LDA is applied on this selected feature space in order to bring it down to a lower dimensional space. It is observed that classification and regression performed on this PCA followed by LDA feature space has increased the accuracy outcome by 8-15% than it is performed on the higher dimensional feature space.

C. Classification

The classification of patient's survival status is done based on eight different categories of patients. The categories are patients who underwent i. Chemotherapy ii. Hormone Therapy iii. Radiotherapy iv. Breast Surgery v. Hormone

Therapy with Breast Surgery vi. Hormone Therapy with Radio Therapy vii. Both Hormone Therapy and Radiotherapy along with Breast Surgery viii. Both Hormone Therapy and Radiotherapy along with Chemotherapy. For each case, the classes are Living and Deceased. The number of samples in each class are mostly equal which makes this dataset balanced.

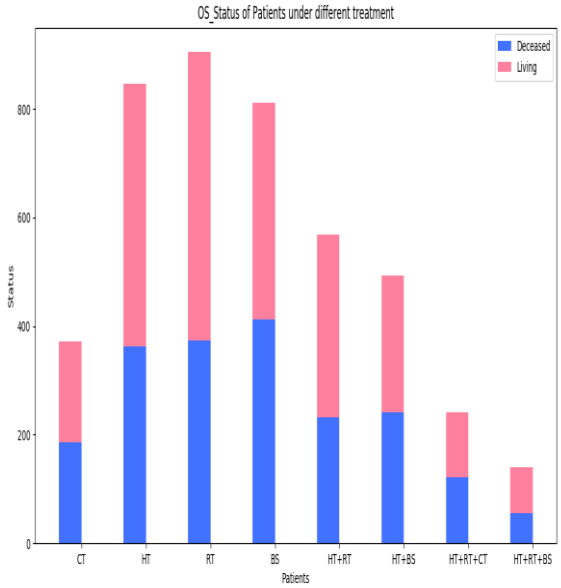

Fig 3: Number of living and deceased patients

Four classifiers are used on each of the classification datasets—Naïve Bayes, SVM with Linear kernel, SVM with RBF kernel and Random Forest Classifier.

1) Naïve Bayes Classifier: Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. It is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable [5]. For example, a fruit may be considered to be an apple if it is red, round, and about 10 cm in diameter. A naive Bayes classifier considers each of these features to contribute independently to the probability that this fruit is an apple, regardless of any possible correlations between the color, roundness, and diameter features.

2) Random Forest Classifier: Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The generalization error for forests converges as to a limit as the number of trees in the forest becomes large [5]. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them. Using a random selection of features to split each node yields error rates that compare favorably to Adaboost (Freund and Schapire [1996]), but are more robust with respect to noise.

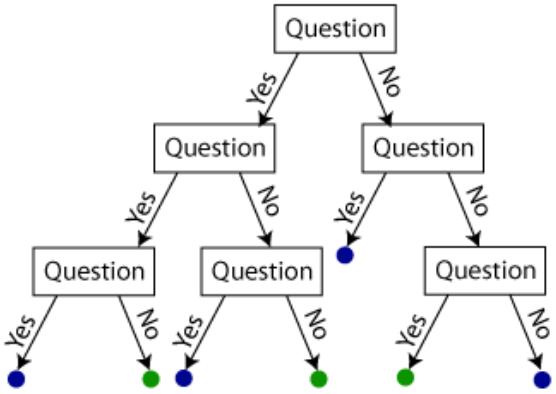

Fig 4: Random Forest Classifier [2]

3) Support Vector Machine: A support vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification. A good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class (functional margin), since in general the larger the margin the lower the generalization error of the classifier.

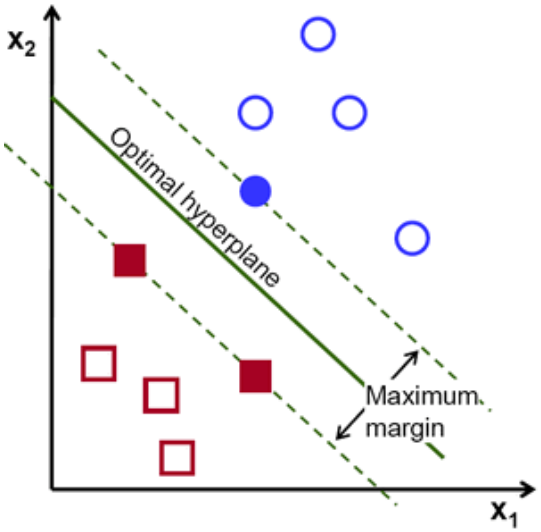

Fig 5: Hyperplane in SVM [3]

In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. An SVM learns to discriminate between the members and non-members of a given functional class based on expression data. Having learned the expression features of the class, the SVM can recognize new genes as members or non-members of the class based on their expression data. In Breast Cancer dataset, linear and RBF kernels are used.

1) Linear Kernel: The application of a support vector machine with a linear kernel is to perform classification or regression. It will perform best when there is a linear decision boundary or a linear fit to the data. In addition, Linear SVM is less prone to overfitting than non-linear. It also works well when the number of features in the given dataset is very large compared to the training samples.

$$Kernel, K(xi,xj) = xi^t . xj, \quad (1)$$

where xi and xj are two vectors.

2) RBF Kernel: Radial Basis Function (RBF) kernel uses Gaussian function to map input space to infinite dimensions. It creates gaussian shaped decision boundary to separate the classes.

$$Kernel, K(xi,xj) = exp\left(-\frac{\|x_i - x_y\|^2}{2\sigma^2}\right), \quad (2)$$

where $\|x_i - x_y\|^2$ is the squared Euclidean distance between the vectors and σ (gamma) controls the shape of the separating function.

D. Regression

In statistical modeling, regression analysis is a set of statistical processes for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables (or 'predictors'). More specifically, regression analysis helps one understand how the typical value of the dependent variable (or 'criterion variable') changes when any one of the independent variables is varied, while the other independent variables are held fixed. In this study, regression task is done to predict NPI score of breast cancer patients based on two types of received treatmments i. Both Hormone Therapy and Radiotherapy along with Breast Surgery ii. Both Hormone Therapy and Radiotherapy along with Chemotherapy and five breast cancer subtypes—i. Basal ii. Claudine-low iii. Her2 iv. LumA v. LumB. The regressor used are Random Forest Regressor, Support Vector Regressor (SVR) with RBF kernel and Kernel Ridge Regressor (KRR) and as an alternative approach Ordinary Least Square (OLS) method is used although it has not fitted well to the datasets.

1) Linear Regression with Ordinary Least Squares: Ordinary least squares (OLS) or linear least squares is a method for estimating the unknown parameters in a linear regression model. OLS chooses the parameters of a linear function of a set of explanatory variables by minimizing the sum of the squares of the differences between the observed dependent variable (values of the variable being predicted) in the given dataset and those predicted by the linear function. Geometrically this is seen as the sum of the squared distances, parallel to the axis of the dependent variable, between each data point in the set and the corresponding point on the regression line – the smaller the differences, the better the model fits the data.

2) Support Vector Regressor: Support vector regressor contains all the main features that characterize maximum margin algorithm: a non-linear function is leaned by linear learning machine mapping into high dimensional kernel induced feature space. The capacity of the system is controlled by parameters that do not depend on the dimensionality of feature space.

3) Kernel Ridge Regressor: Kernel ridge regression (KRR) combines ridge regression (linear least squares with l2-norm regularization) with the kernel trick. It thus learns a linear function in the space induced by the respective kernel and the data. For non-linear kernels, this corresponds to a non-linear function in the original space. The form of the model learned by KRR is identical to support vector regression (SVR). However, different loss functions are used: KRR uses squared error loss while support vector regression uses epsilon-insensitive loss, both combined with l2 regularization. In contrast to SVR, fitting a KRR model can be done in closed-form and is typically faster for medium-sized datasets. On the other hand, the learned model is non-sparse and thus slower than SVR, which learns a sparse model for epsilon > 0, at prediction-time.

4) Random Forest Regressor: Random forests can be used for regression analysis and are in fact called Regression Forests. They are an ensemble of different regression trees and are used for nonlinear multiple regression. Each leaf contains a distribution for the continuous output variables. in a regression tree, since the target variable is a real valued number, a regression model is fitted to the target variable using each of the independent variables. Then for each independent variable, the data is split at several split points. The Sum of Squared Error(SSE) is calculated at each split point between the predicted value and the actual values. The variable resulting in minimum SSE is selected for the node. Then this process is recursively continued till the entire data is covered.

E. Cross-validation and Model Selection

Evaluating the performance of classifiers and regressors and selecting the best model out of many are two most crucial part of any machine learning study. For this research, 10-fold cross-validation method is used for classification perfromace evaluation.

1) 10-fold cross-validation: 10-fold cross validation splits the entire dataset into 10 disjoint sets of equal size. The remaining samples are distributed over some of the sets. In this method, the classifier is trained 10 times. Each time one of the ten sets are removed and the classifier is trained on the remaining nine sets. Then the model is evaluated on the separated test set. After each iteration, the accuracy measures are stored for different test sets. End of the iterations 10 accuracy scores are obtained for ten different sets of test data. Finally, the mean of all the scores are taken into account in order to attain the ultimate accuracy of the model.

2) Model Selection: For each fold, one ROC curve is generated and the area under the curve (AUC) is calculated. The model that gives the highest AUC with the lowest number of predictors is considered as the best model for any individual

dataset. On the other hand, for regression, the model that gives out the smallest Sum of Squared Error (SSE) is considered as the best model for that individual dataset.

IV. RESULT AND DISCUSSION

A. *Classification Experiments*

Classification is done to classify the survival status of eight categories of patients based on the treatment they received. The aim of this experiment is to reduce the number of predictors while increasing the level of accuracy. After applying all the four classifiers on each of the cases, there performance is compared by the individual AUC.

In Fig 6, the accuracy of classifying patient survival status who solely received Chemotherapy is 82% for Random Forest, 88% for SVMLin, 89% for SVMRBF and 87% for Naïve Bayes.

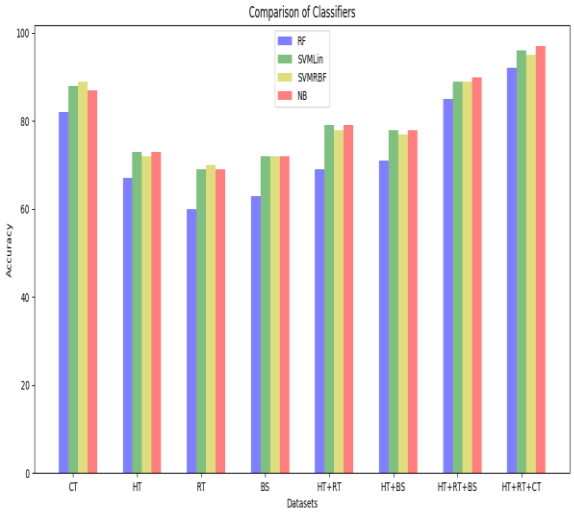

Fig 6: Performance comparison between different classifiers

On the other hand, the accuracy of classifying patient survival status who solely received Hormone therapy is only 67% for Random Forest, 73% for SVMLin, 72% for SVMRBF and 73% for Naïve Bayes. The accuracy of classifying patient survival status receiving only Radio therapy is 60% for Random Forest, 69% for SVMLin, 70% for SVMRBF and 69% for Naïve Bayes. However, survival classification of patients underwent Mastectomy breast surgery is obtained with accuracy of 63% for Random Forest, 72% for SVMLin, 72% for SVMRBF and 72% for Naïve Bayes. If the combination of therapies is taken into consideration, then accuracy level is increased for all the classifiers. As example, the accuracy of classifying patient survival status who received both Hormone Therapy and Radio is 69% for Random Forest, 79% for SVMLin, 78% for SVMRBF and 79% for Naïve Bayes. On the other hand, the accuracy of classifying patient survival status who received both Hormone Therapy and Breast Surgery is 71% for Random Forest, 78% for SVMLin, 77% for SVMRBF and 78% for Naïve Bayes. If the combination of three treatments are considered, then the survival status of patients becomes even more easy to classify as the classes become more well separated. As example, the accuracy of classifying patient survival status who received all three of treatments—Hormone Therapy, Radio Therapy and Breast Cancer is 85% for Random Forest, 89% for SVMLin, 89% for SVMRBF and 90% for Naïve Bayes. In addition, the accuracy of classifying patient survival status who received all three of treatments—Hormone Therapy, Radio Therapy and Chemotherapy is 92% for Random Forest, 96% for SVMLin, 95% for SVMRBF and 97% for Naïve Bayes. From the discussion above, it is evident that SVM Lin, SVM RBF and Naïve Bayes are able to successfully classify the survival status for all the eight cases whereas Random Forest is not giving a satisfactory outcome even after using equal number of features for the all the classifiers.

TABLE I. MODEL A = LASSO + ELASCTICNET + RFE

Datasets	Model A			
	No. of Selected Genes	AUC	No. of Selected Genes	AUC
CT	25	0.79	50	0.83
HT	28	0.69	50	0.79
RT	18	0.68	50	0.72
BS	**27**	**0.72**	50	0.73
HT+RT	25	0.76	50	0.79
HT+BS	**27**	**0.78**	50	0.82
HT+RT+BS	27	0.88	50	0.93
HT+RT+CT	20	0.93	50	1.0

TABLE II. MODEL B = ELASCTICNET + RFE

Datasets	Model B			
	No. of Selected Genes	AUC	No. of Selected Genes	AUC
CT	**25**	**0.88**	50	0.91
HT	**28**	**0.75**	50	0.80
RT	**18**	**0.70**	50	0.73
BS	27	0.37	50	0.48
HT+RT	**25**	**0.79**	50	0.80
HT+BS	**27**	**0.78**	50	0.82
HT+RT+BS	27	**0.89**	50	0.92
HT+RT+CT	20	**0.97**	50	0.99

Table I and Table II compare the performance of the two models based on individual AUC obtaining from using equal number of genes. According the tables, in seven out of eight cases Model B is giving better accuracy by using both lower number of features and top 50 features than Model A. Only in case of Breast Surgery, Model A outperforms Model B as it is giving 72% accuracy using 27 genes and 73% accuracy using 50 genes whereas Model B is giving only 37% and 48% accuracy respectively. Although, increasing the number of features increase the overall accuracy, classify with more features require more computational time and cost. Therefore, when there is no significance improvement in accuracy between 50 genes and a lower number of genes, the latter is selected as a model.

Finally, Model A with respective number of selected genes is used for the Breast Surgery dataset and Model B with respective number of selected genes is used for the rest of the seven datasets.

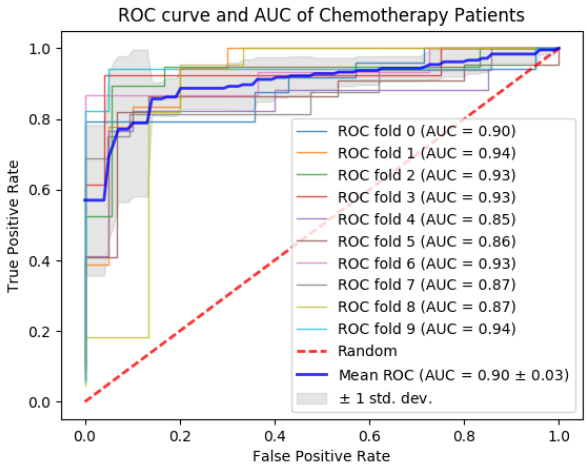

Fig 7: 10-fold ROC and AUC for Chemotherapy Patients

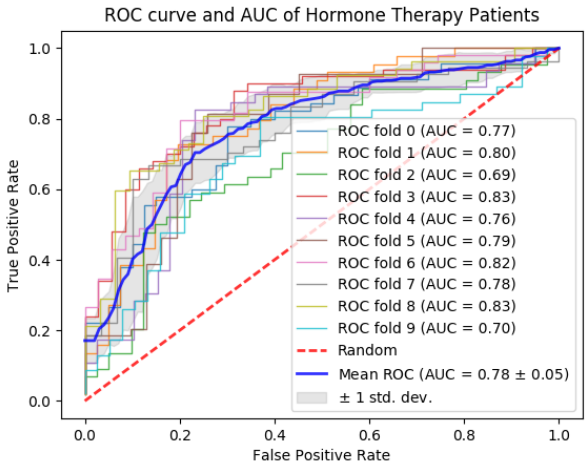

Fig 8:10-fold ROC and AUC for Hormone Therapy Patients

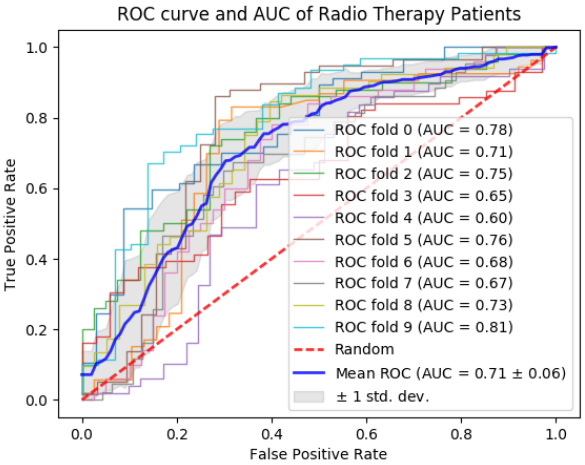

Fig 9: 10-fold ROC and AUC for Radio Therapy Patients

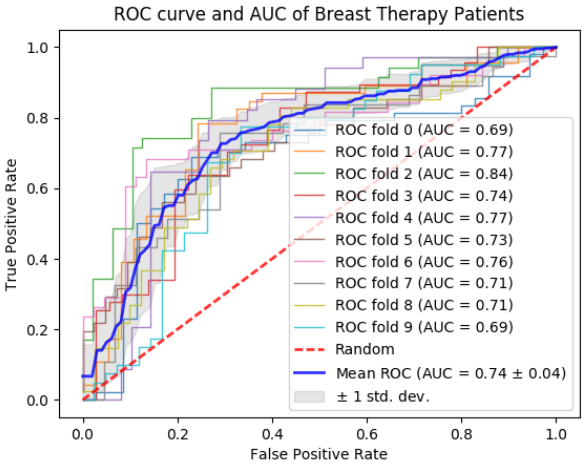

Fig 10: 10-fold ROC and AUC for Breast Surgery Patients

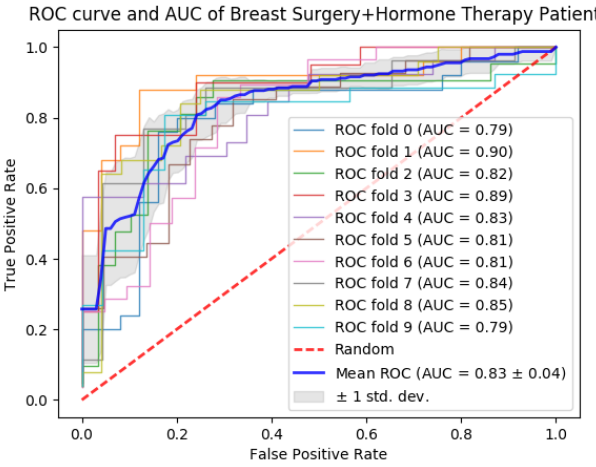

Fig 11:10-fold ROC and AUC for Breast Surgery + Hormone Therapy Patients

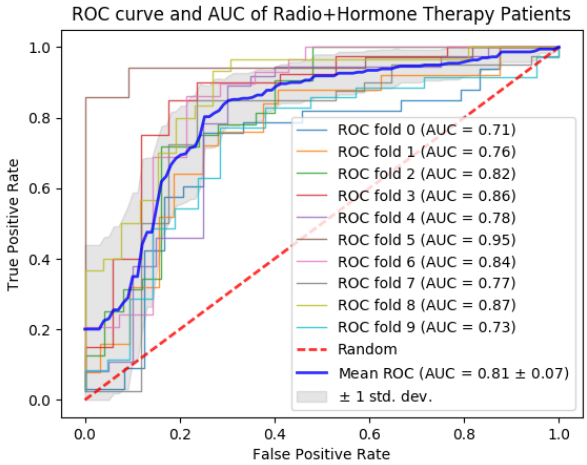

Fig 12: 10-fold ROC and AUC for Radio + Hormone therapy Patients

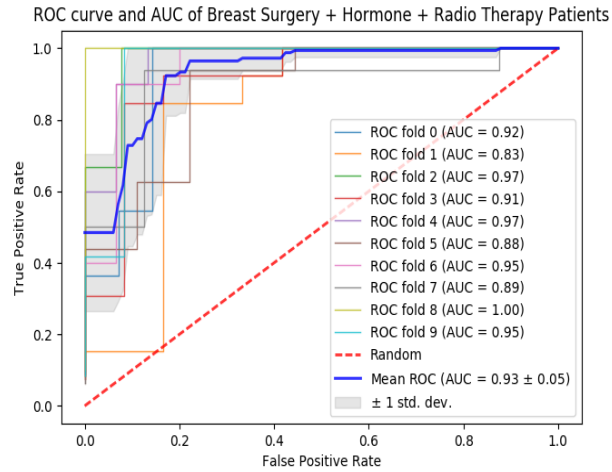

Fig 13:10-fold ROC and AUC for Breast Surgery + Hormone + Radio therapy Patients

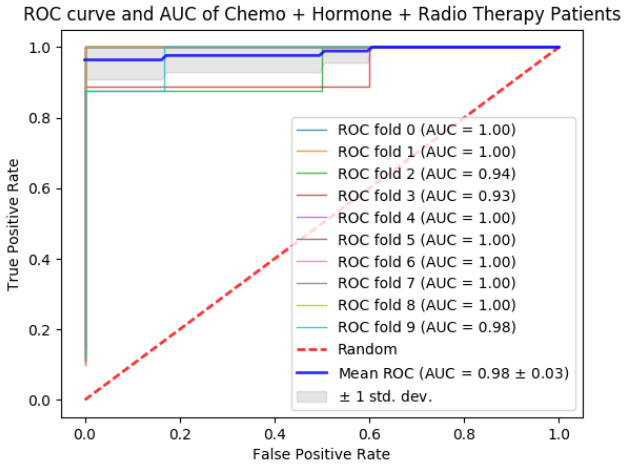

Fig 14:10-fold ROC and AUC for Chemo + Hormone + Radio therapy Patients

Fig. 7-14 shows the ROC curve and corresponding area under the curve for all the 10-folds for all the eight cases using the best feature selection model with SVMLin Classifier. The mean ROC and AUC is obtained by taking the average of accuracies from all the 10 folds. For each case, individual standard deviation and confidence interval is also calculated. For example, in Fig 14, the mean AUC is 0.98 with a standard deviation of 0.03.

B. Regression Analysis

Regression analysis is done to predict NPI scores on two categories of patients based on the treatment they received – i. Both Hormone Therapy and Radiotherapy along with Breast Surgery ii. Both Hormone Therapy and Radiotherapy along with Chemotherapy and five breast cancer subtypes—i. Basal ii. Claudine-low iii. Her2 iv. LumA v. LumB. The regressor used are Random Forest Regressor, Support Vector Regressor (SVR) with RBF kernel and Kernel Ridge Regressor (KRR) and as an alternative approach Ordinary Least Square method is used although it has not fitted well to the datasets.

To predict the NPI scores of patients who underwent three combinations of treatments, Random Forest, OLS, SVR and KRR are fitted. In all the cases, the number of selected features is 23. The regression plots are generated for applying the fitted model on held-out test dataset.

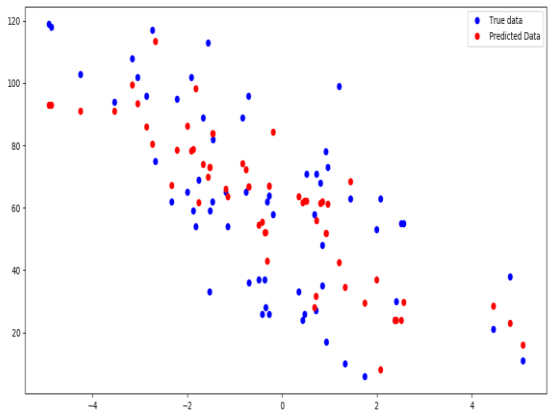

Fig 15: Random Forest Regressor on dataset (i)

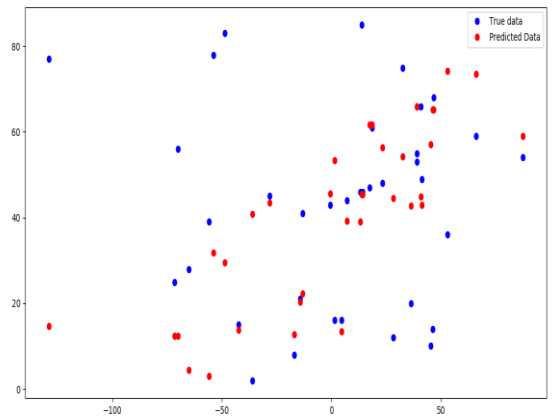

Fig 16: Random Forest Regressor on dataset (ii)

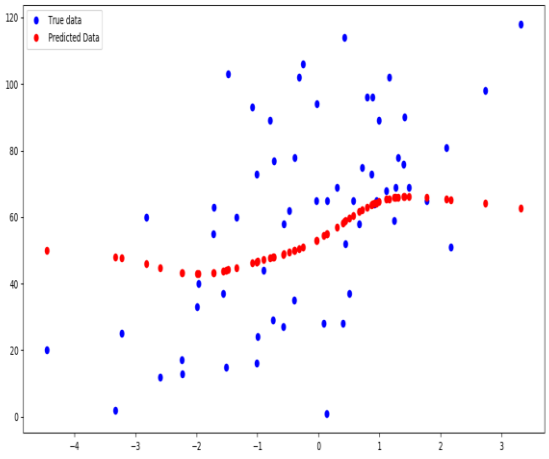

Fig 17: Support Vector Regressor on dataset (i)

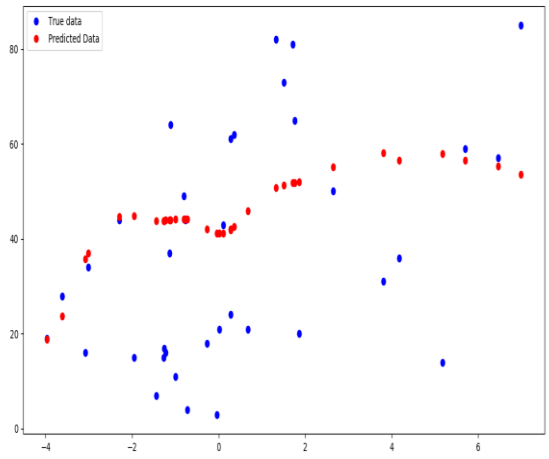

Fig 18: Support Vector Regressor on dataset (ii)

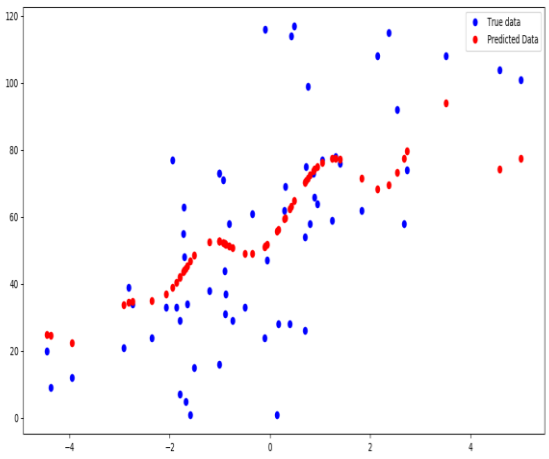

Fig 19: Kernel Ridge Regressor on dataset (i)

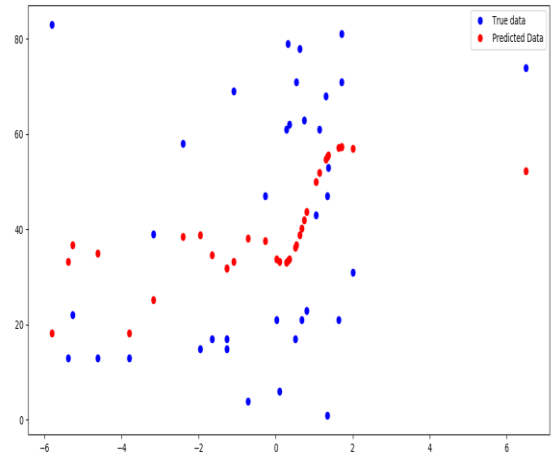

Fig 20: Kernel Ridge Regressor on dataset (ii)

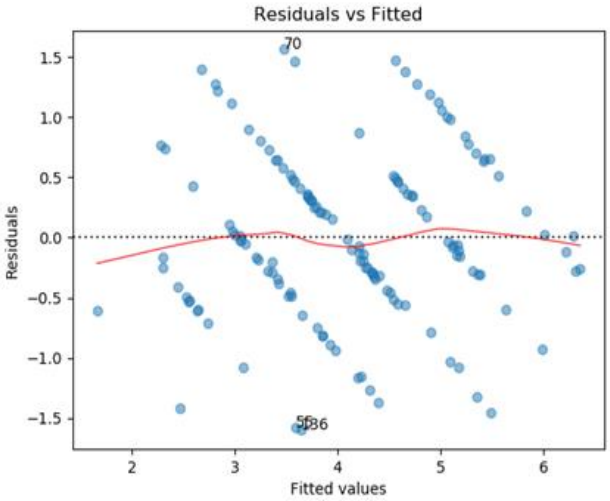

Figure 21: Residual vs Fitted for OLS for dataset (i)

In Fig. 15 and 16, Random Forest Regressor is fitted on dataset (i) and (ii). In dataset (i), RFR is fitted quite well to the true data whereas in dataset (ii), it is fitted to some extent although due to the outliers the sum of squared error will be increased.

In Fig. 17 and 18, Support Vector Regressor is fitted on dataset (i) and (ii). In dataset (i), SVR is not fitted well to the true data. In dataset (ii), it is fitted better than dataset (i) however still it is not fitted in a satisfactory manner.

In Fig. 19 and 20, Kernel Ridge Regressor is fitted on dataset (i) and (ii). In dataset (i), KRR is not fitted well to the true data. In dataset (ii), it is fitted better than dataset (i) however still it is not fitted in a satisfactory manner.

In Fig 21, Ordinary Least Squares method(OLS) is used on dataset (i) and the linear regression model using OLS is fit to the training dataset. The result of R-squared for this dataset is 0.913. R-squared is the percentage of variance explained by a model.

According to the regression plots, it can be deduced that RFR is fitted to the training dataset more accurately compared to both SVR and KRR. SVR performs the poorest on both of the datasets. On the other hand, a high R-squared value of OLS in dataset (i) indicates that the model is very well explained by the fitted model.

To predict the NPI scores of patients based on five cancer subtypes, SVR and KRR are fitted. In all the cases, the number of selected features is 23. The regression plots are generated for applying the fitted model on held-out test dataset.

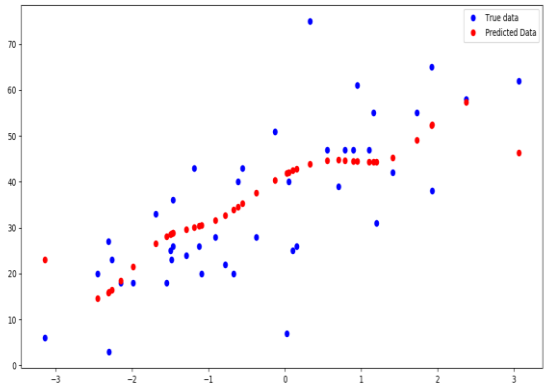

Fig 22: KRR on Basal subtype

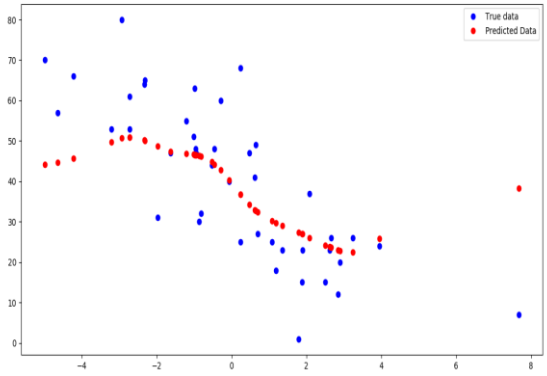

Fig 23: SVR on Basal Subtype

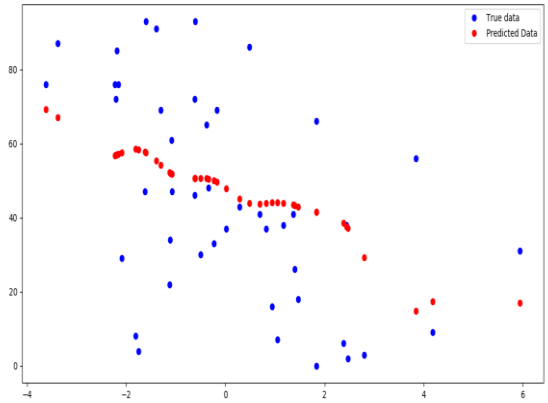

Fig 24: KRR on Claudin-low subtype

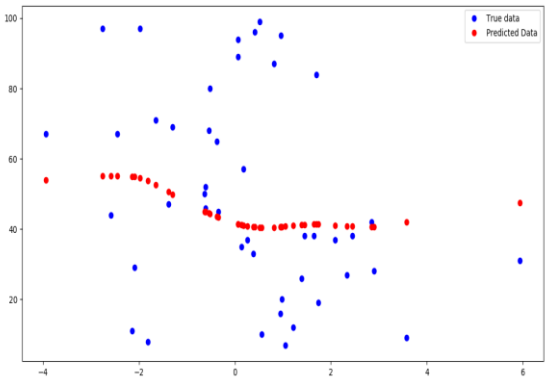

Fig 24: SVR on Claudin-low subtype

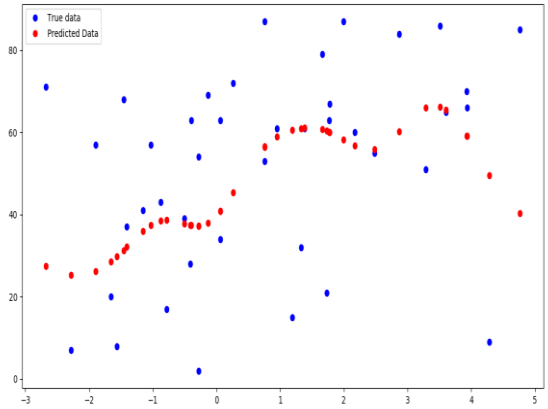

Fig 25: KRR on Her2 subtype

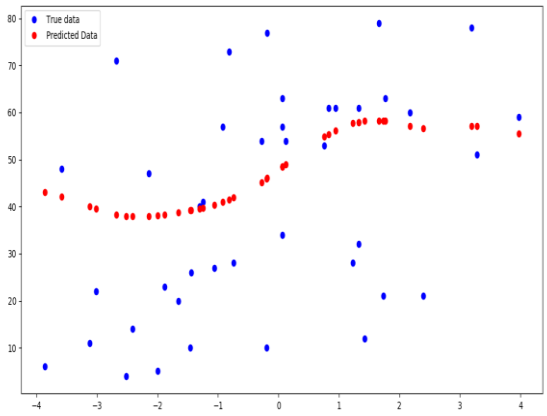

Fig 26: SVR on Her2 subtype

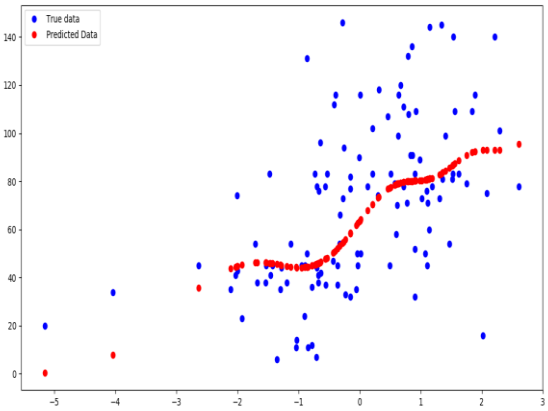

Fig 27: KRR on LumA subtype

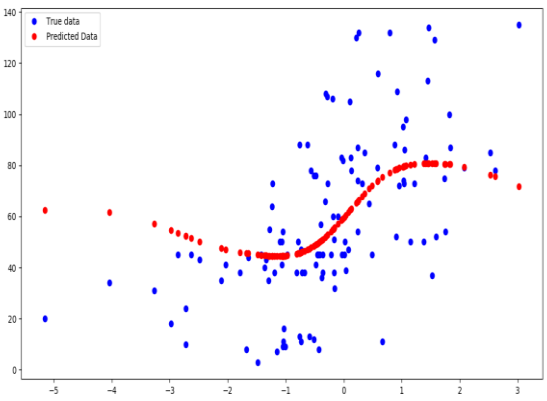

Fig 28: SVR on LumA subtype

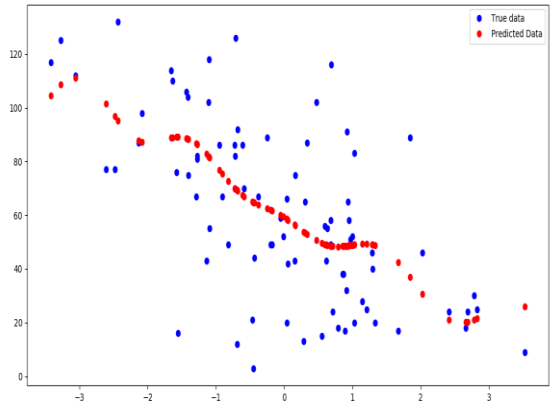

Fig 29: KRR on LumB subtype

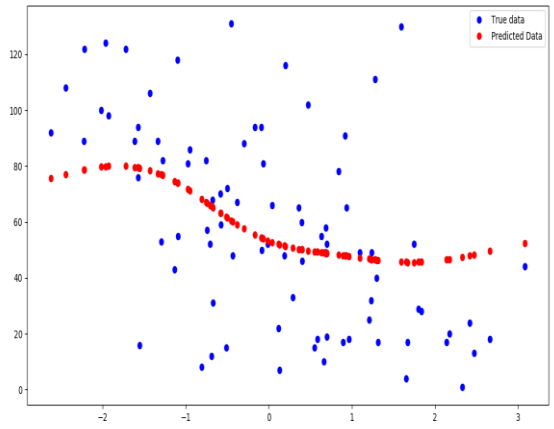

Fig 30: SVR on LumB subtype

In Fig. 21 and 22, KRR and SVR is fitted on Basal dataset. According to the regression plots, KRR is fitted better than SVR in this dataset.

In Fig. 23 and 24, KRR and SVR is fitted on Claudin-low dataset. According to the regression plots, both KRR and SVR haven't fitted well to this dataset.

In Fig. 25 and 26, KRR and SVR is fitted on Her2 dataset. According to the regression plots, KRR is fitted slightly better than SVR in this dataset.

In Fig. 27 and 28, KRR and SVR is fitted on LumA dataset. According to the regression plots, both KRR and SVR performed almost similar in fitting to this dataset.

In Fig. 29 and 30, KRR and SVR is fitted on LumB dataset. According to the regression plots, KRR is fitted better than SVR in this dataset.

According to the regression plots, it can be deduced that KRR is fitted to the training dataset more accurately compared to both SVR.

V. Future Work

As future work, multi-layer perceptron based neural network can be implemented to sole the above performed tasks.

Multilayer perceptron (MLP) is applied in many fields due to its powerful and stable learning algorithm. In some previous studies [7], it is used to perform different classification tasks in bioinformatics with success. Applying the same method in the classification and prediction problems explained here can be an extension in future.

VI. CONCLUSION

In this study, combination of different classification, regression and feature selection methods are used to solve some of the most challenging classification and prediction tasks in bioinformatics. The obtained accuracy is satisfactory and also gained from using very limited number of predictors or genes. Proposed methods are applied to eight classification and seven regression datasets of breast cancer patients. Model performance is evaluated using 10-fold cross-validation methods applied on separate test datasets. It is reported that the choice of feature selection methods, the number of genes in the gene list, the number of cases (samples) substantially influence classification success. Based on features chosen by the proposed methods, AUC and accuracy of several classification algorithms are obtained. Results reveals the importance of feature selection in accurately classifying new samples and how an integrated feature selection and classification algorithm is performing and is capable of identifying significant genes.

REFERENCES

[1] What is breast cancer? - Canadian Cancer Society. (n.d.). Retrieved December 22, 2017, from http://www.cancer.ca/en/cancer-information/cancer-type/breast/breast-cancer/?region=on

[2] Random forests. (2013, July 10). Retrieved December 22, 2017, from https://shapeofdata.wordpress.com/2013/07/09/random-forests/

[3] Introduction to Support Vector Machines. (n.d.). Retrieved December 22, 2017, from https://docs.opencv.org/2.4.13.4/doc/tutorials/ml/introduction_to_svm/introduction_to_svm.html

[4] Feature Selection of Gene Expression Data for Cancer Classification: A Review. (2015, May 08). Retrieved December 22, 2017, from http://www.sciencedirect.com/science/article/pii/S187705091500561X?via%3Dihub

[5] González-Navarro, F. F., Belanche-Muñoz, L. A., & Silva-Colón, K. A. (n.d.). Effective Classification and Gene Expression Profiling for the Facioscapulohumeral Muscular Dystrophy. Retrieved December 22, 2017, from http://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0082071

[6] Bouazza, S. H., Hamdi, N., Zeroual, A., & Auhmani, K. (2015). Gene-expression-based cancer classification through feature selection with KNN and SVM classifiers. *2015 Intelligent Systems and Computer Vision (ISCV)*. doi:10.1109/isacv.2015.7106168

[7] Dash, S., & Dash, A. (2014). A correlation based multilayer perceptron algorithm for cancer classification with gene-expression dataset. *2014 14th International Conference on Hybrid Intelligent Systems*. doi:10.1109/his.2014.7086190