Classification Models and Survival Analysis for Prostate Cancer Using RNA Sequencing and Clinical Data

Md Faisal Kabir
Department of Computer Science
North Dakota State University, NDSU
Fargo, USA
mdfaisal.kabir@ndsu.edu

Simone A. Ludwig
Department of Computer Science
North Dakota State University, NDSU
Fargo, USA
simone.ludwig@ndsu.edu

***Abstract*—Early detection of cancer can significantly increase the chance of successful treatment. This research performs a study on early cancer detection for prostate cancer patients from whom cancer tissue was analyzed with Illumina Hi-Seq ribonucleic acid (RNA) Sequencing (RNA-Seq). Cancer relevant genes with the most significant correlations with the clinical outcome of the sample type (cancer / non-cancer) and the overall survival (OS) were assessed. Traditional cancer diagnosis primarily depends on physicians' experience to identify morphological abnormalities. Gene expression level data can assist physicians in detecting cancer cases at a much earlier stage and thus can significantly improve the potential of patient treatment. In this research, for the classification task, we applied machine learning and data mining approaches to detect cancer versus non-cancer based on gene expression data. Our goal was to detect cancer at the earliest stage. Besides, for the regression task, survival outcomes in prostate cancer patients were performed. Regression trees were built using cancer-sensitive genes along with clinical attribute 'Gleason score' as predictors, and the clinical variable 'overall survival' as the target variable. Knowledge in the form of rules is one of the vital tasks in data mining as it provides concise statements of easily understandable and potentially valuable information. For the classification model, we derived rules from a decision tree and interpreted these rules for cancer and non-cancer patients. For the regression or survival model, we generated rules for predicting or estimating the survival time of cancer patients. In this study, cancer-relevant genes were analyzed as predictors, although various genes may interact with genes currently known to contribute to cancer. These findings have implications for assessing gene-gene interactions and gene-environment interactions of prostate cancer as well as for other types of cancer.**

***Index Terms*—data mining, RNA Sequencing, Genomic data Commons, Prostate Cancer, rules generation, Survival Analysis.**

I. Introduction

Cancer has become one of the most devastating diseases worldwide, with more than 18.1 million new cases every year [1]. Cancer incidence and mortality are rapidly growing worldwide. Prostate cancer is the second most common malignancy among men worldwide after lung cancer, with $1,276,106$ new cases causing 358,989 deaths (3.8% of all deaths caused by cancer in men) in 2018 [1]. While incidences of prostate cancer are high, about 95 % of all prostate cancers can be detected when the disease is limited to the prostate. Also, treatment success rates are high compared to most other types of cancer. Generally, the earlier the cancer is caught and treated, the more likely the patient will remain disease-free. For that reason, it is crucial to detect prostate cancer cases as early as possible.

Molecular signatures hold the promise of precise and systemic cancer diagnosis and classification. The cellular activity is dynamically regulated during pathological conditions through changes in the expression level in genes. Therefore, specific gene expression profiles are necessary signatures that are helpful for early diagnosis because gene expression abnormalities always appear before morphological changes can be observed [2]. Authors in [3] examined if the three randomly selected cancer-related genes were correlated during cancer progression and whether they showed the association between gene expressions and early cancer development for breast, kidney, liver, and thyroid cancers. Therefore, our goal is to build a classifier using machine learning or data mining techniques to detect early cancer cases based on gene expression data.

Personalized medicine may be promoted by assessing genomic and clinical variables simultaneously. Genes may interact with one another and clinical variables are overall survival, cancer stages, sex, and so forth. Early detection of cancer leads to an increase in survival rate, and consideration of clinical variables along with RNA-Seq data may be utilized to increase efforts at the early detection of cancer.

In this research, for detecting cancer different classification models were built and for the prediction or estimation of survival time several regression models were studied. Gene expression of prostate data of cancer-relevant genes along with the clinical variable 'Gleason score' were used as predictors. For the classification task, sample type (cancer / non-cancer) was used as the target variable while for regression or survival prediction the overall survival (OS) was used as the target variable.

This paper is comprised of five important sections following the introduction. The related work is discussed in Section II. In

Section III, the methodology of our research is elaborated on including data characteristics, feature selection techniques for both classification and survival analysis. Also, model building along with brief descriptions of the algorithms are provided. The experiments and results are illustrated in Section IV. The results obtained from various feature selection models are provided and discussed. Also, the rule generation from both decision tree and regression tree are shown in this section. Section V is our discussion section. Section VI is the summary section of this paper where we conclude our paper and suggest possible future research directions.

II. Related Work

In machine learning or data mining, classification is an example of supervised learning techniques. The goal of classification training a classification model is to predict qualitative or categorical outputs which assume values in a finite set of classes without an explicit order [4]. Regression models are used to predict one variable from one or more variables. Regression learns a function that maps a data item to a real-valued prediction variable. Many regression methods exist in mathematics, such as linear, non-linear, logistic, and multi-linear regression. Regression models provide the data miner with a powerful tool, allowing predictions about past, present, or future events to be made with information about past or present events [4]. Data mining is often referred to as knowledge discovery in databases and describes the process of nontrivial extraction of implicit, previously unknown and potentially valuable information from a large amount of data [5]. The mined information is referred to as knowledge provided in the form of rules, constraints, and regularities. In data mining, rule mining is one of the vital tasks since rules provide concise statements of potentially valuable information that can be easily understood by end users [6].

Researchers have developed different statistical, data mining, and machine learning models for various cancers detection and estimation of survival time. In most of the cases, the researchers used clinical or patient data. However, gene expression abnormalities always appear before morphological changes can be observed. Therefore, in this research we build the model by investigating gene expression data along with the clinical data.

Next-generation sequencing has revolutionized the field by not only increasing the sequencing depths and accuracy but also reducing the time and cost to an affordable level for individual cancer patients. Therefore, gene expression profiling has become a feasible cancer diagnosis and prognosis. Researchers have developed various models with promising results. Authors in [7] investigated six different machine learning techniques on publicly available datasets of predicting cancer outcome. Besides, the authors also used different feature selection approaches of identifying relevant genes for maximizing predictive information. In [8], the early diagnosis of breast cancer is done using genetic algorithms (GA) along with artificial neural networks (ANN). The authors used GA for feature extraction and parameter optimization of the ANN. Rule generation is one of the vital tasks since rules provide concise statements of potentially relevant information that can be easily understood by end users [6]. The authors in [9], discovered useful rules of breast cancer and non-breast cancer patients from risk factors data using association rule mining techniques.

In this research, we used the gene expression of prostate data of cancer-relevant genes along with a clinical variable 'Gleason score' as predictors. For the classification task, sample type (cancer / non-cancer) was used as the target variable, while for regression or survival prediction the overall survival (OS) was used as the target variable. Furthermore, knowledge in the form of rules was generated from both the classification and regression models. These rules can be useful for physicians or biologists to investigate i) the relationship between the overall survival and specific gene expression levels, and ii) the association between sample type and specific gene expression levels in prostate cancer.

III. Methods

A. Data Characteristics

RNA-seq and clinical variables available from the National Cancer Institute Genomic Data Commons (GDC) were investigated in this research. These variables were integrated in order to detect cancer cases and survival predictions based on the level of individual variables as well as the interaction of these variables, including RNA-Seq and clinical predictors. Illumina Hi-Seq RNA sequencing $log2(x + 1)$ normalized data was merged with clinical variables accessible from the GDC.

There were a total of 550 instances in the prostate cancer data set. Among them, 497 were primary tumor samples (cancer patients), and 52 were solid tissue standard samples (non-cancer individuals). There was only one sample named as metastatic tumor, which has been considered as a primary tumor. So, total primary tumor or cancer samples counted were 498, and normal or non-cancer samples were 52. There were more than twenty thousand (20,000) genes, however, in this research we only consider 36 common genes that are associated with cancer (according to the National Cancer Institute Genomic Data Commons [10]).

Thirty-six (36) cancer-relevant genes (AR, BRCA1, BRCA2, CD82, CDH1, CHEK2, EHBP1, ELAC2, EP300, EPHB2, EZH2, FGFR2, FGFR4, GNMT, HNF1B, HOXB13, IGF2, ITGA6, KLF6, LRP2, MAD1L1, MED12, MSBM, MSR1, MXI1, NBN, PCNT, PLXNB1, PTEN, RNASEL, SRD5A2, STAT3, TGFBR1, WRN, WT1, and ZFHX3) and clinical variable 'Gleason score' (an index of cancer stage) of prostate cancer were assessed as predictors of tissue type (cancer or non-cancer).

Besides, these cancer-sensitive genes, along with clinical variable 'Gleason score' of prostate cancer were assessed as predictors in survival analysis to predict overall survival (OS). The goal of the survival analysis is to increase the ability to predict survival time based on the expression level of predictors genes and the clinical variable 'Gleason score'. The distribution of overall survival is shown in Fig. 1.

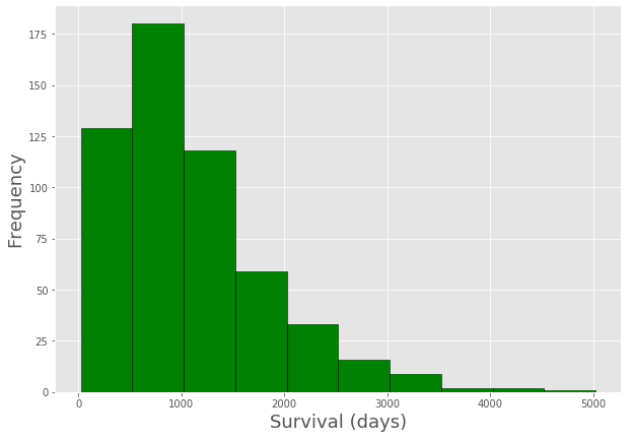

Fig. 1. Distributions of clinical variable overall survival (OS).

B. Feature Selection Approaches

A multivariate correlation analysis was performed to observe the correlation of predictor variables with the target variable. Predictors were filtered and sorted with the absolute correlation coefficient value. A value closer to 0 implies a weaker relationship, and a value closer to 1 means a stronger correlation with the target. For the classification model, a multivariate correlation analysis was performed to observe the association of predictor variables with the target variable named as sample type (cancer or non-cancer). For survival prediction, the correlation was done with the target variable named as overall survival (OS). Predictors or genes were filtered and sorted with the absolute correlation coefficient value with cancer/sample type and then with OS, respectively.

The area of feature selection in machine learning has become quite robust. There are numerous feature selection algorithms which identify the features from given data that contributes the most to the target variable [11]. An extra-trees classifier and select-K-best approaches were investigated to obtain relevant or essential features for building the classification models. An extra-tree or extremely randomized trees classifier [12] implements a meta-estimator that fits several randomized decision trees named as extra-trees on various sub-samples of the data set. It is very similar to a Random Forest Classifier and only differs in the way the construction of the decision trees is done using the forest. In the feature selection process, the Gini index is used in the creation of the forest. Each feature is ordered in descending order according to the Gini importance of each feature, and the user can select the top K features accordingly. The Select-K-best algorithm extracts features according to the highest scores. It calculates a chi-square statistic between each feature and the target variable. The implementation of these algorithms was performed using the Scikit-learn python package [13].

The Cox (proportional hazards or PH) model is the most commonly used multivariate approach for analyzing survival time data in medical research [14]. The Cox regression model extends survival analysis methods to assess the effect of several risk factors on survival time simultaneously. The model is used to identify the impact of predictors on the survival of cancer patients. This model makes it possible to isolate variables that have little effect on survival. Furthermore, the model allows estimating the risk or danger of death for an individual based on the prognostic (good for survival) variables. The output of the Cox (ph) regression model, along with the hazard ratio, was investigated to select a good predictor (good prognostic factor) for survival. The Hazard Ratio (hr) assesses the overall survival or the risk of death by the predictors. Generally, the value of hazard ratio less than 1.0 is considered a good predictor (good prognostic factor) for survival, while the value of hr greater than 1.0 is considered not good for survival (bad prognostic factor).

C. Classification and Regression Techniques

The classification techniques that we investigated in this paper are decision tree (DT), random forest (RF), and multi-layered neural network (MLP or NN). Besides, for the survival analysis, the decision tree regressor was investigated.

Decision tree induction is the learning of decision trees from class-labeled training tuples. A decision tree is a flowchart-like tree structure, where each internal node (non-leaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label [15] [16]. It works very well with different types of data, and results are easy to understand. The process of model building is comparatively easy compared to other algorithms, and data can be represented in a visual form (tree-like form). From the tree, we can generate or form rules that can be used to classify unknown values. The decision tree classifier has been widely applied to solve many real-world problems in different fields [17] [18].

Random forest is a robust classification and regression technique that generates a forest of classification trees, rather than a single classification tree. RF creates trees on randomly selected data instances and obtains the prediction from each tree to choose the best solution through voting. RF is considered as a highly accurate and robust technique because it generates many trees in the process [18] [19].

A neural network is a set of connected input/output units in which each connection has a weight associated with it [4]. During the learning phase, the network learns by adjusting the weights to be able to predict the correct class label of the input tuples. Neural networks involve long training times and are therefore more suitable for applications where this is feasible. The most popular neural network algorithm is back-propagation – Multilayer feed-forward networks. A multilayer feed-forward neural network consists of an input layer, one or more hidden layers, and an output layer.

A regression tree is similar to a classification tree, except that the target variable is continuous, and a regression model is fitted to each node to return the prediction value of target variable [20]. Here, the tree is used to predict the value for unknown cases. For regression, the prediction error is typically

measured by the squared difference between the observed and predicted values.

D. Building Models

For cancer detection, we built classifier models with 36 cancer-sensitive genes and the clinical variable 'Gleason score'. The target variable is tissue or sample type (cancer or non-cancer). A decision tree, a random forest, and multi-layer neural networks were selected as classifiers. The default parameter values were used for the random forest algorithm. For the decision tree algorithm, the maximum depth of the tree was specified as six, and for the multi-layered neural network two hidden layers with 25 and 12 nodes were used. The same procedure was followed with predictors or genes that were considered or selected using the feature selection approaches.

For building the classification models and the prediction (survival) model, we split the data into 70 % training set, and 30 % test set with stratified train test split.

E. Rule Generation from Tree

From the built trees, we generated knowledge in the form of rules. For the classification model, a decision tree was built, and from the tree, rules were generated for both cancer and non-cancer patients. For the regression model, we created rules for the estimation of survival time. To obtain a rule, we need to follow the tree down from the root to the leaf nodes.

IV. Experiments and Results

Results of the feature extraction for both the classification models and survival prediction are discussed in this section. Moreover, the performance measure of the classifiers and both classification (decision) tree and regression tree are shown here. Finally, knowledge discovery in the form of rules from both decision tree (cancer detection) and regression tree (survival prediction) are shown and elaborated.

A. Output of Feature Selection for Classification Model

Genes correlated with sample or cancer type were determined. Correlations of selected cancer-relevant genes with a clinical variable named as sample type were represented in heat maps and genes in the order of those with the highest absolute value of association with cancer type are EZH2, HOXB13, RNASEL, FGFR2, SRD5A2, CD82, MXI1, MAD1L1, IGF2, ITGA6, PTEN.

The important genes with clinical variable sample type that were obtained using the extra tree classifier are shown as a bar graph in Fig. 2. The genes are given in the order of the importance, which are EZH2, FGFR2, HOXB13, CD82, RNASEL, SRD5A2, MAD1L1, PCNT, MSMB, WRN, WT1, LRP2, MXI1, FGFR4, PLXNB1.

The SelectKBest technique selects K best features according to the highest scores. Fifteen (K = 15) predictors or genes according to the highest score are: SRD5A2, FGFR2, EZH2, LRP2, HOXB13, IGF2, CD82, WT1, RNASEL, HNF1B, GNMT, PTEN, EPHB2, KLF6, MSR1 are shown in Fig. 3.

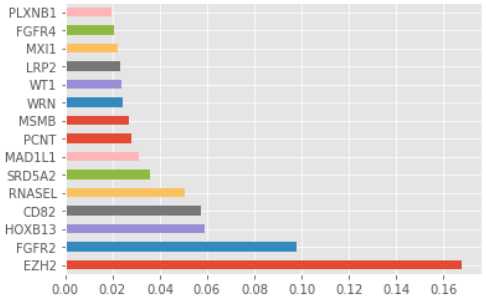

Fig. 2. Important features that were obtained using Extra Tree Classifier.

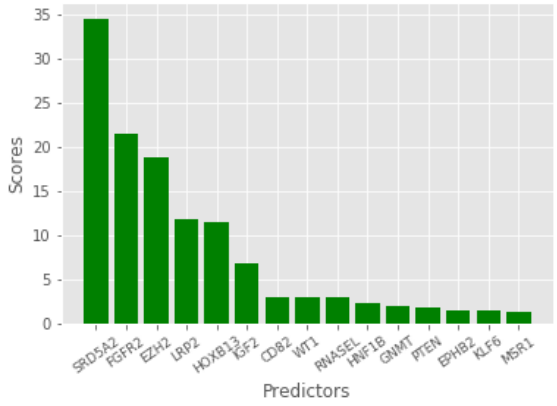

Fig. 3. Feature selection using K best features.

B. Selected Predictors for Classification Model

The three aforementioned feature extraction approaches were applied. Most essential predictors that were common in all three techniques are EZH2, HOXB13, RNASEL, FGFR2, SRD5A2, and CD82. We also selected more features (MXI1, MAD1L1, IGF2, PTEN, WT1, and LRP2) that were common in any two techniques.

C. Performance Measure of Classifiers

To evaluate the performance, several measures were used such as accuracy, recall, precision, area under the Receiver Operating Characteristic curve (ROC) or AUC, and F-measure [18] [19] [21]. These were derived from the confusion matrix and applied to the classifier evaluation.

$$Accuracy = (TP + TN)/(TP + FP + TN + FN) \quad (1)$$

$$recall = TP/(TP + FN) \quad (2)$$

$$precision = TP/(TP + FP) \quad (3)$$

Here, TP denotes the number of positive examples correctly classified, TN denotes the number of negative samples correctly classified, FN represents the number of positive observations incorrectly classified, and FP indicates the number of negative samples incorrectly classified by the estimator. The ROC curve is a representation of the best decision boundaries for the cost between the True Positive Rate (TPR) and the

False Positive Rate (FPR). The ROC curve plots TPR against FPR. TPR and FPR are defined as follows:

$$TPR = TP/(TP + FN) \tag{4}$$

$$FPR = FP/(FP + TN) \tag{5}$$

The F-measure or F1 score is a measure of a test's accuracy and is defined as the weighted harmonic mean of the precision and recall of the test, which is defined as follows:

$$F - measure = 2 * \frac{precision * recall}{precision + recall} \tag{6}$$

Detailed information about these measures can be found in [18] [19].

D. Results of Classifiers

In this paper, we applied three different classifier models on the training data and compared the performance of the trained models on the test data. The overall performance of the classification models is shown in Table I. Results were evaluated on the test data.

TABLE I
OVERALL PERFORMANCE BASED ON TEST DATA.

Methods	Precision	Recall	F1-score	Accuracy (%)	AUC
DT	0.92	0.92	0.92	92.1212	0.7611
RF	0.96	0.96	0.96	95.7576	0.8920
MLP	0.94	0.94	0.94	93.9393	0.9421

Area under the Receiver Operating Characteristic curve (ROC) or AUC for these three classifiers are shown in Fig. 4.

In the second step of our classification technique, we trained data that were obtained using feature selection approaches (discussed in Section IV-B). The overall performance of classifiers are shown in Table II.

TABLE II
OVERALL PERFORMANCE BASED ON TEST DATA (CLASSIFIERS TRAINED WITH SELECTED FEATURES).

Methods	Precision	Recall	F1-score	Accuracy (%)	AUC
DT	0.91	0.91	0.91	90.9090	0.89052
RF	0.96	0.93	0.93	93.3333	0.8744
MLP	0.93	0.93	0.93	93.9393	0.90772

Comparing both tables, we can see that in general multi-layered neural networks (MLP) performs better when we trained the model without the feature selection approach. If we look at the F1 measure, which is the weighted harmonic mean of the precision and recall, the classifiers trained with all features (without feature selection) perform well compared to the trained models with the selected predictors (important features). The reason for this is that all the features contribute to the detection of prostate cancer rather than using fewer predictors.

E. Generated Rules from Decision Tree

In Fig. 5, a tree was shown that was built by applying the decision tree classifier with all 36 genes and the Gleason score as predictors. The root node, with the most information gain indicates the significant gene in determining cancer or non-cancer for prostate data, which is EZH2. The impurity is the measure as given at the top by the Gini score. Samples show the number of instances available to classify, and the value indicates how many samples are in class 0 (non-cancer) and how many samples are in class 1 (cancer).

If we follow the tree down from the root to the leaf nodes, we can find a rule. From the tree, we generated some rules for both cancer and non-cancer patients that are shown as follows:

Cancer patients:

Rule 1: If the gene expression level of EZH2 is less than or equal 5.494, and the gene expression level of CD82 is less or equal 9.51, then there is a chance that individual will be a cancer patient.

Rule 2: If the gene expression level of EZH2 is less than or equal 5.494 and the gene expression level of CD82 is higher than 9.51, and the gene expression value for CHEK2 is less or equal 5.567, and EHBP1 is less or equal 9.982 then there is a chance that individual will be a cancer patient.

Rule 3: If the gene expression level of EZH2 is less than or equal 5.494 and the gene expression level of CD82 is higher than 9.51, and the gene expression value for CHEK2 is greater than 5.567, and TGFBR1 is less or equal 9.605 then there is a chance that individual will be a cancer patient.

Rule 4: If the gene expression level of EZH2 is less than or equal 5.494 and the gene expression level of CD82 is higher than 9.51, and the gene expression value for CHEK2 is more than 5.567, and TGFBR1 is more than 9.605, and MED12 is less or equal 10.765 then there is a high chance that individual will be a cancer patient.

Rule 5: If the gene expression level of EZH2 is less than or equal 5.494 and the gene expression level of CD82 is higher than 9.51 and the gene expression value for CHEK2 is more than 5.567 and TGFBR1 is more than 9.605 and MED12 is higher than 10.765, and the gene expression level of EZH2 is less than 3.918 then there is a chance that individual will be a cancer patient.

Non-Cancer patients:

Rule 1: If the gene expression level of EZH2 is less than or equal 5.494 and the gene expression level of CD82 is higher than 9.51, and the gene expression value for CHEK2 is less or equal to 5.567, and EHBP1 is more than 9.982 then an individual will not be a cancer patient.

Rule 2: If the gene expression level of EZH2 is less than or equal 5.494 and the gene expression level of CD82 is higher

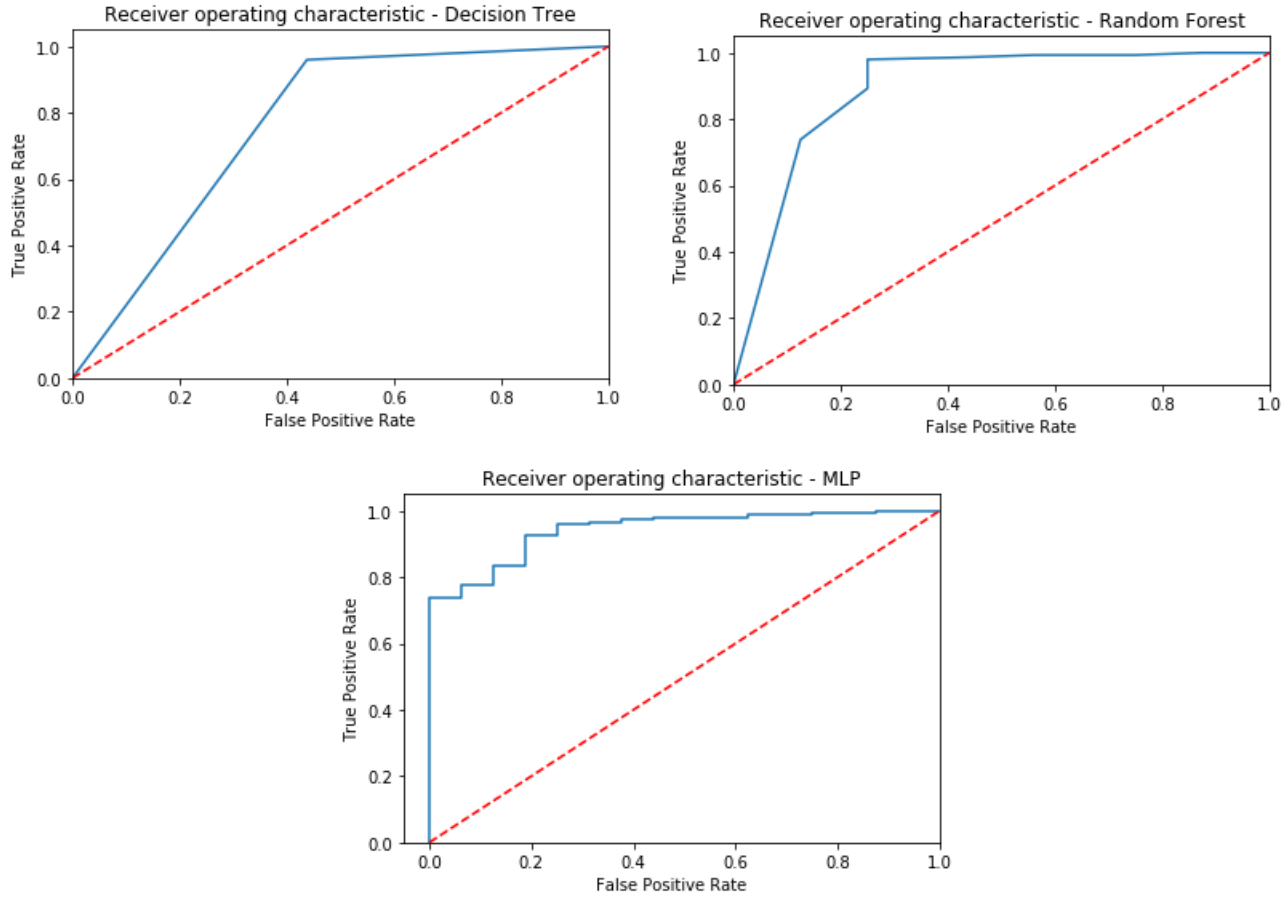

Fig. 4. ROC curve for three specified classifiers.

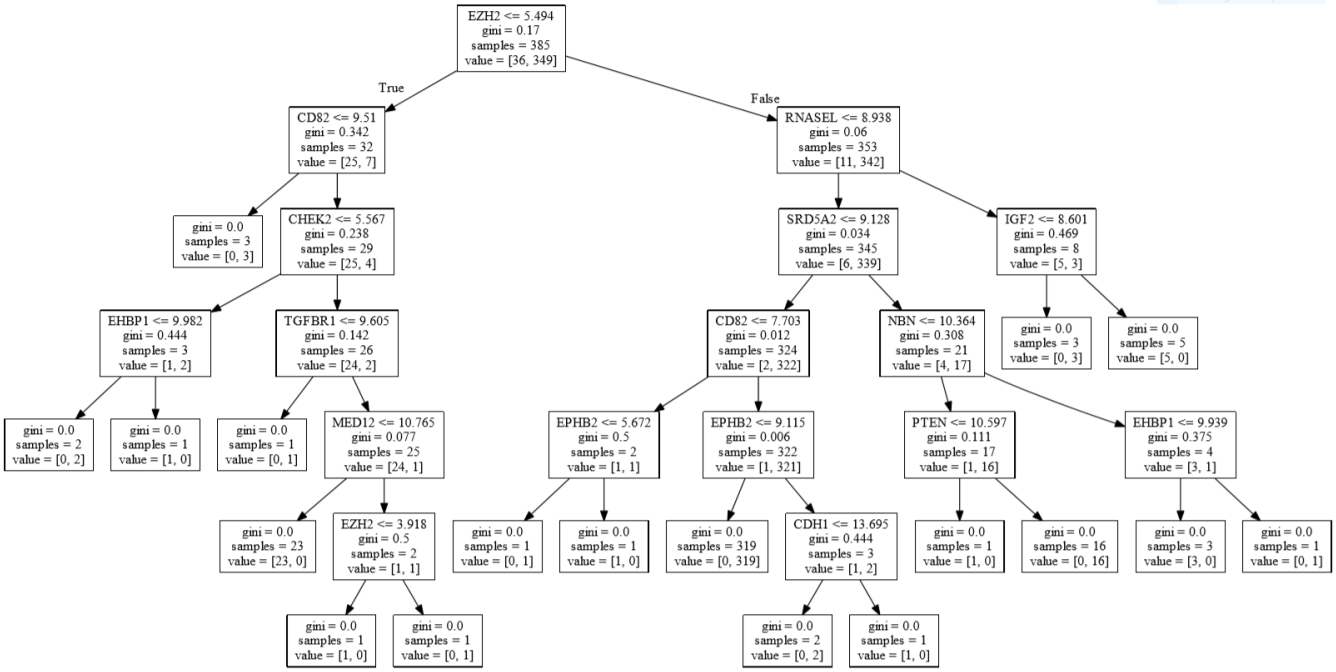

Fig. 5. A decision tree that was built by using cancer-sensitive genes (without feature selection).

than 9.51 and the gene expression value for CHEK2 is greater than 5.567 and TGFBR1 is more than 9.605 and MED12 is higher than 10.765, and the gene expression level of EZH2 is more than 3.918 then there is a chance that individual will be a non-cancer patient.

Rule 3: If the gene expression level of EZH2 is higher than 5.494 and the gene expression level of RNASEL is more than 8.938, and the gene expression value for IGF2 is less than 8.601 then there is a chance that individual will be a non-cancer patient.

F. Results of Feature Selection for Survival Prediction

Genes correlated with a clinical variable overall survival (OS) were determined. Correlations of selected genes with clinical variable overall survival are represented in a heat map that is shown in Fig. 6. Genes are given in the order of those with the greatest absolute value of correlation with overall survival (OS): AR, BRCA2, CD82 , CDH1, EPHB2, FGFR2, FGFR4, IGF2, ITGA6, LRP2, MAD1L1, MED12, MSMB, MSR1, PLXNB1, RNASEL, ZFHX3.

In the Cox (ph) regression model, the p-value for all three tests likelihood ratio test (p = 0.008), Wald test (p= 0.02), and Score (log-rank) test (p= 0.02) are significant, indicating that the model is significant. Also, in the multivariate Cox analysis, the covariates BRCA1, EZH2, and MED12 remain significant. However, other covariates fail to be significant. The output of the Cox (ph) regression model along with the hazard ratio are shown in Fig. 7. The hazard ratio (HR) assesses the overall survival or the risk of death by predictors. Good predictors or good prognostic factors that were obtained by applying multivariate Cox (proportional hazards) regression based on hazard ration are BRCA1, CHEK2, EHBP1, EP300, EPHB2, GNMT, HNF1B, IGF2, ITGA6, MAD1L1, MSR1, MXI1, NBN, PCNT, PLXNB1, SRD5A2, WRN, gleason_score.

G. Decision Tree Regressor for Survival Predictions

We build three regression models by applying the decision tree regressor. In the first model, all cancer sensitive genes along with the clinical variable gleason_score were used as the predictor for predicting the survival time (overall survival - OS). In the second model, variables that had a higher correlation with the overall survival were considered. In the final model, variables that were obtained from the Cox (ph) regression model based on the hazard ratio were used for predictors. For the performance evaluation, mean square error (MSE) was considered for the test data. Among these three models, the second model was selected for further study as it has a lower MSE value than the other models.

In Fig. 8, a tree was shown that was built by applying the decision tree regressor on higher correlation genes with overall survival (OS). The root node can be considered as the most informative feature or gene for survival prediction. In our cases, MED12 is the most informative gene and then LRP2 or BRCA2 based on the expression value of MED12.

Predictions of Survival Time from Decision Tree Regressor:

The root node MED12 can be considered as the most important gene for overall survival prediction. If we visit from the root node to a particular leaf node, we can find a rule for survival time prediction. From the regression tree, we can generate the number of rules or knowledge that will be helpful to predict patients' survival time. Some of the rules generated from the regression tree are shown as follows:

Rule 1: If the gene expression level of MED12 is less than or equal 8.818 and the gene expression level of LRP2 is less or equal 1.247 and expression level of CD82 is less or equal to 9.142, and the gene expression level of ITGA6 is less or equal to 10.307 then there is a chance that the patient will survive about 3502 days.

Rule 2: If the gene expression level of MED12 is less than or equal 8.818 and the gene expression level of LRP2 is less or equal 1.247 and expression level of CD82 is less or equal to 9.142, and the gene expression level of ITGA6 is higher than 10.307 then there is a chance that the patient will survive about 3440 days.

Rule 3: If the gene expression level of MED12 is less than or equal 8.818 and the gene expression level of LRP2 is less or equal 1.247 and expression level of CD82 is higher than 9.142 then there is a likelihood that the patient will survive about 2850 days.

Rule 4: If the gene expression level of MED12 is higher than 8.818, and the gene expression level of BRCA2 is less or equal 0.363, then there is a chance that the patient will survive about 4264 days.

Rule 5: If the gene expression level of MED12 is higher than 8.818 and the gene expression level of BRCA2 is more than 0.363 and IGF2 is less or equal 5.238 then there is a possibility that the patient will survive about 3467 days.

Rule 6: If the gene expression level of MED12 is higher than 8.818, and the gene expression level of BRCA2 is more than 0.363, and IGF2 is greater than 5.238, and the gene expression level of FGFR4 is more abundant than 6.644 and MSR1 is less or equal to 7.161 then there is a possibility that the patient will survive about 971 days.

Rule 7: If the gene expression level of MED12 is higher than 8.818 and the gene expression level of BRCA2 is greater than 0.363 and IGF2 is larger than 5.238 and the gene expression level of FGFR4 is greater than 6.644 and MSR1 is higher than 7.161 then there is a chance that the patient will survive about 1682 days.

From the regressor tree, we can generate rules as discussed above and can estimate or predict the survival time or overall

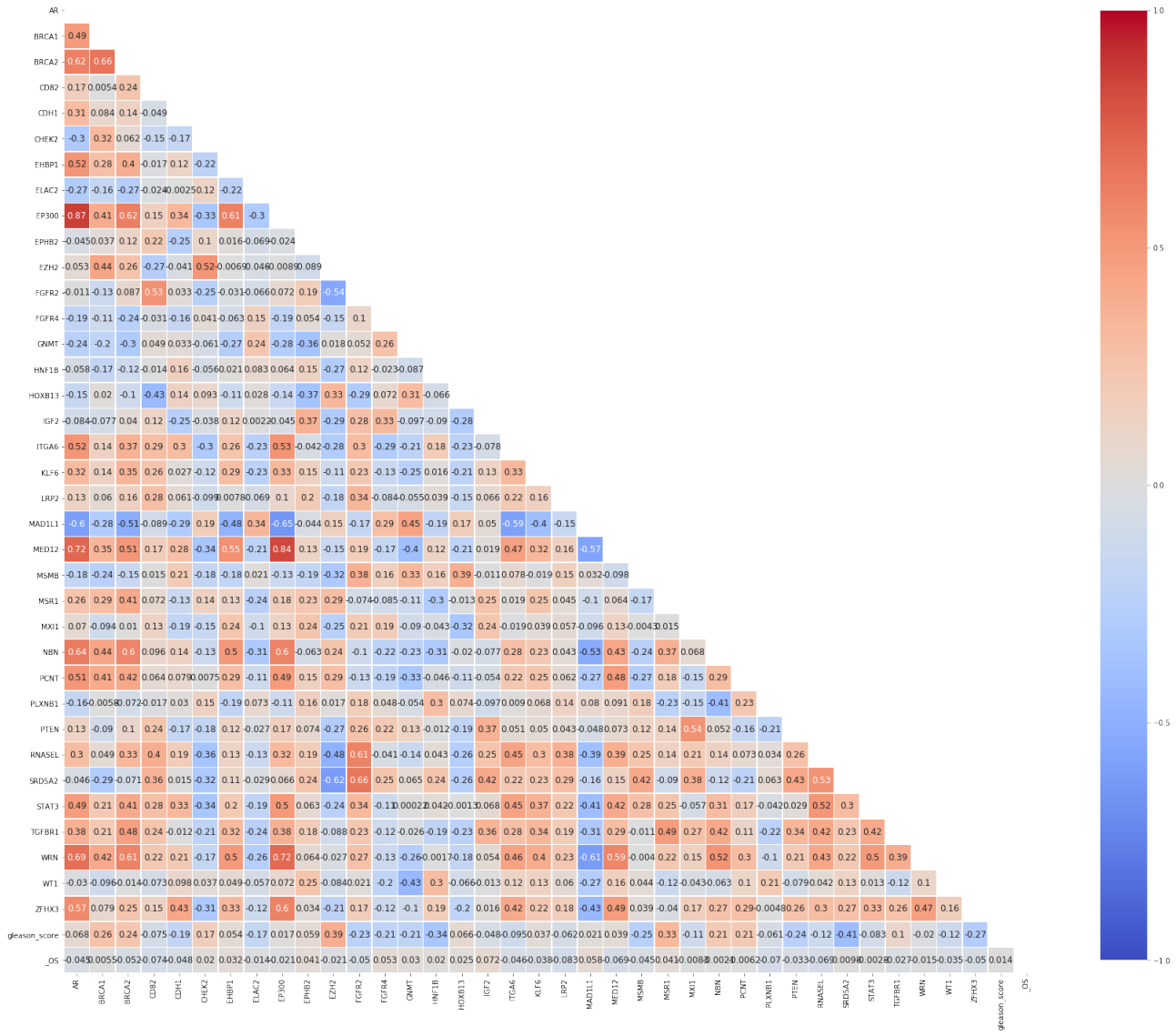

Fig. 6. Heat map of correlations of cancer-relevant genes with a clinical variable overall survival (OS).

survival for a particular patient.

V. Discussion

The National Institutes of Health Genomic Data Commons may be utilized to determine which clinical variables and RNA-Seq expression levels detect clinical outcomes, such as sample types and overall survival. In this research, in order to get a clear understanding of RNA-Sequencing and clinical data, we investigated 36 cancer-sensitive genes and few clinical variables. Based on the classification models for cancer detection, we see that the model performs better for unseen cases when we applied all 36 genes and the clinical variable 'Gleason score' as predictors; instead of applying only a few predictors (obtained by using feature selection approaches). This has implications that for predicting cancer cases, almost all features contribute rather than the selected features. It also implies that for building classification models in cancer detection, all genes (about twenty-thousand) along with other clinical variables should be investigated further.

Furthermore, in survival prediction or estimation, we see the model that uses higher correlated features with overall survival (OS) performs better than the other models. Overall, the correlation of features with overall survival (OS) was very low, which also implies that all genes contribute to the overall survival. This means that for our further studies in survival prediction we should use all the predictors.

In this research, we also generated rules from the decision tree and the regression tree. By looking at the rules, we can see that the level of gene expression plays a vital role in determining if an individual could be a cancer patient or non-cancer patient. For instance, have a look at the rule (Fig. 5 - part of the right subtree), if the level of expression of the gene

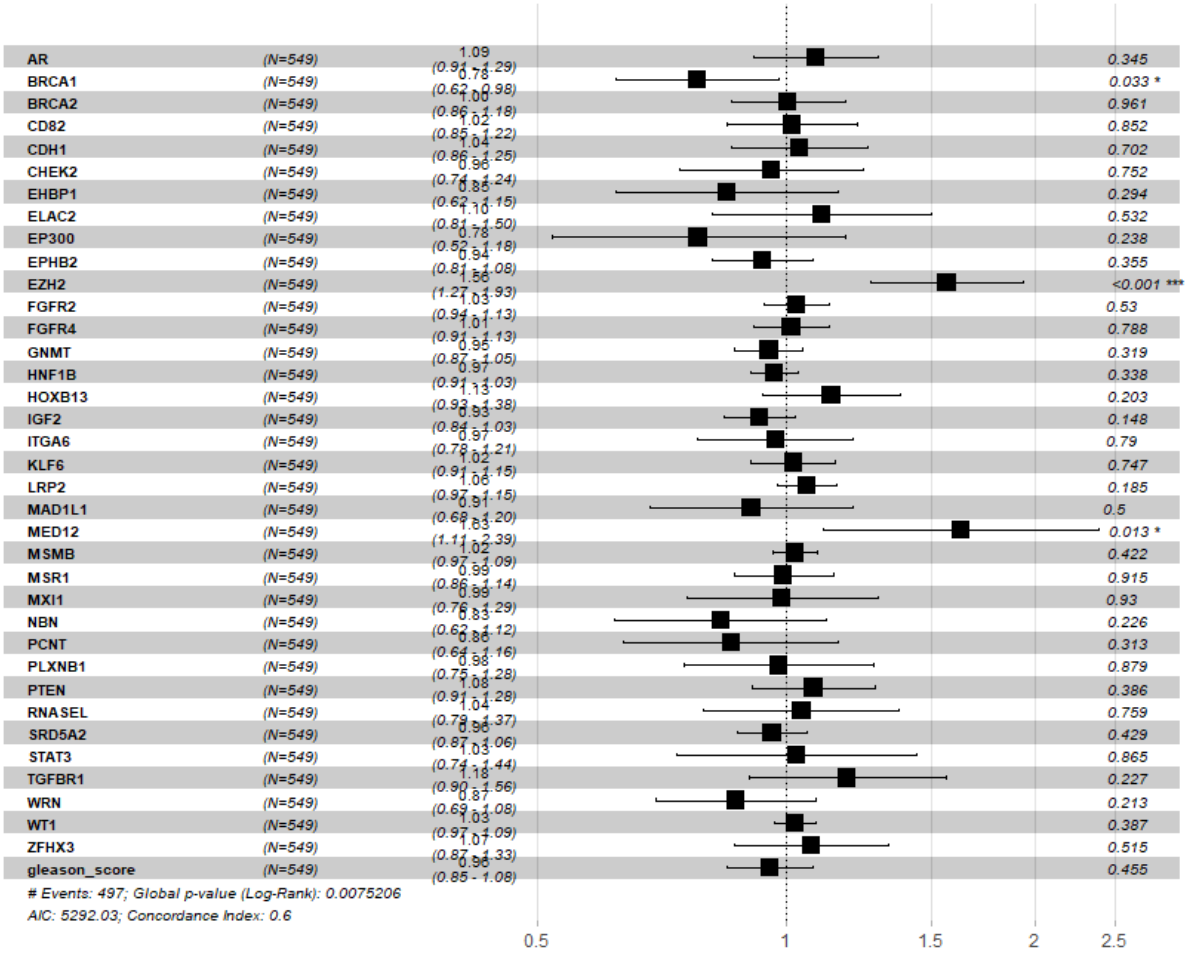

Fig. 7. The output of Cox (ph) regression model along with hazard ratio.

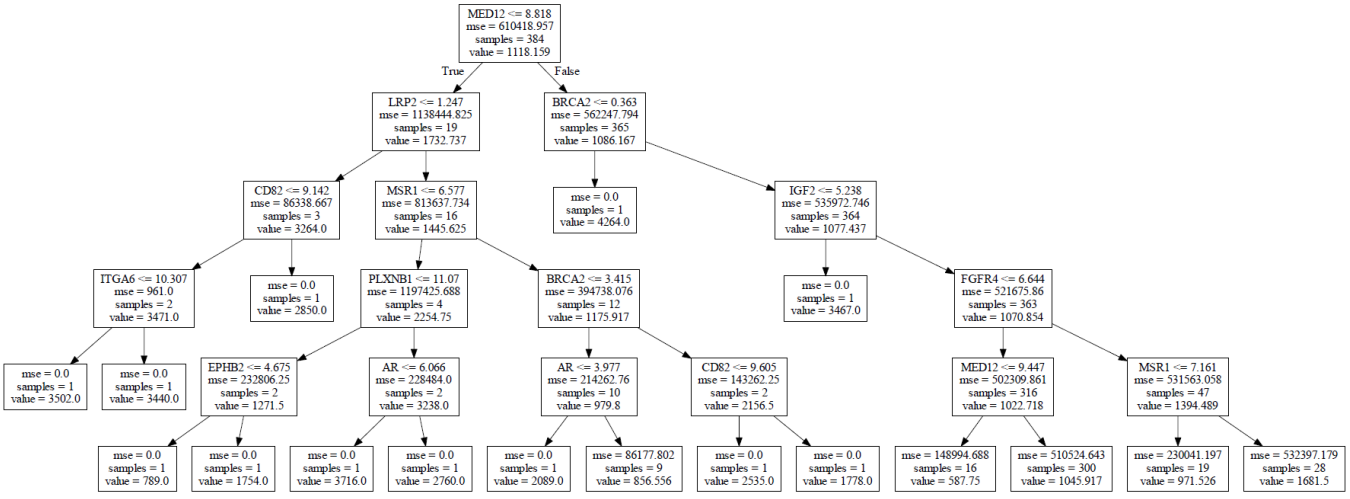

Fig. 8. A regressor tree that was built by using higher correlation genes with overall survival (OS).

EZH2 is greater than 5.494, and the expression level of gene RNASEL is larger than 8.938 then the gene expression level of IGF2 plays a key function in determining cancer or non-cancer for a particular patient. If the gene expression level IGF2 is less or equal to 8.601, then there is a possibility that an individual will not have cancer; otherwise, there is a high chance that the patient will have cancer. These types of relationships among various genes with corresponding expression levels and clinical variables can be further investigated for personalized medicine research. These type of associations can be found from the regression tree as well.

VI. Conclusion

RNA-seq and clinical variables available on the National Cancer Institute Genomic Data Commons (GDC) were investigated in this research. For detecting clinical variable cancer type, we built three different classification models based on decision tree (DT), random forest (RF), and multi-layered neural networks (MLP) using gene expression data. Different feature selection techniques were also investigated to find the most predictive genes, and we developed models using the three aforementioned classification methods on these selected genes. The results showed that MLP performs better on test data when we built the model without applying any feature selection approach.

Also, the prediction of the clinical variable 'overall survival' in prostate cancer was performed by applying i) all 36 genes and the clinical variable 'Gleason score' as predictors, and ii) genes obtained from the feature selection approach. Furthermore, rule generation was performed from a selected decision tree classifier for both cancer and non-cancer patients. Rules discovery was also performed from a selected regression tree for estimating survival outcome.

In this research, we utilized 36 cancer-sensitive genes along with few clinical variables. Future studies will assess all genes (about twenty-thousand) along with more clinical variables.

Acknowledgment

The data set that was used for this investigation is the one that was provided by the 3rd IEEE Big Data Governance and Metadata Management (http://ieeesa.io/bdgmm) workshop/hackathon under the theme of Big Data Bioscience Data Mashup and Analytics that was held during the IEEE Big Data conference in Seattle, USA, on December 10-13, 2018. The first author of this paper is one of the 1st place winners of Hackathon Track 1 on Personalized Medicine for Drug Targeting in Prostate Cancer Patients.

References

[1] Bray, Freddie, et al. "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries." CA: a cancer journal for clinicians 68.6 (2018): 394-424.

[2] Okamoto, Oswaldo Keith. "DNA microarrays in cancer diagnosis and prognosis." Einstein 3.1 (2005): 31-34.

[3] Ling, Binbing, et al. "Gene expression correlation for cancer diagnosis: a pilot study." BioMed research international 2014 (2014).

[4] Rahman, SM Monzurur, Md Faisal Kabir, and Muhammad Mushfiqur Rahman. Integrated Data Mining and Business Intelligence. Encyclopedia of Business Analytics and Optimization. IGI Global, 2014. 1234-1253.

[5] J. Han, M. Kamber. "Data mining concept and technology." Publishing House of Mechanism Industry: 70-72, 2001.

[6] S. M. Monzurur Rahman, Md. F. Kabir, and F. A. Siddiky. "Rules mining from multi-layered neural networks." International Journal of Computational Systems Engineering1.1: 13-24, 2012.

[7] Luque-Baena, Rafael Marcos, et al. "Application of genetic algorithms and constructive neural networks for the analysis of microarray cancer data." Theoretical Biology and Medical Modelling 11.1 (2014): S7.

[8] Ahmad, Fadzil, et al. "Intelligent breast cancer diagnosis using hybrid GA-ANN." 2013 Fifth International Conference on Computational Intelligence, Communication Systems and Networks. IEEE, 2013.

[9] Kabir, Md Faisal, Simone A. Ludwig, and Abu Saleh Abdullah. "Rule Discovery from Breast Cancer Risk Factors using Association Rule Mining." 2018 IEEE International Conference on Big Data (Big Data). IEEE, 2018.

[10] GDC. GDC, 2018, portal.gdc.cancer.gov/, accessed on January, 2019.

[11] Hemphill, Edward, et al. "Feature selection and classifier performance on diverse bio-logical datasets." BMC bioinformatics. Vol. 15. No. 13. BioMed Central, 2014.

[12] Geurts, Pierre, Damien Ernst, and Louis Wehenkel. "Extremely randomized trees." Machine learning 63.1 (2006): 3-42.

[13] Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." Journal of machine learning research 12.Oct (2011): 2825-2830.

[14] Husain, Hartina, et al. "The Application of Extended Cox Proportional Hazard Method for Estimating Survival Time of Breast Cancer." Journal of Physics: Conference Series. Vol. 979. No. 1. IOP Publishing, 2018.

[15] J. R. Quinlan, Constructing decision tree. C4 5, 17-26, 1993.

[16] Y. Zheng, et al. R-C4. 5 Decision tree model and its applications to health care dataset. Services Systems and Services Management, 005. Proceedings of ICSSSM05. 2005 International Conference on. Vol. 2. IEEE, 2005.

[17] Md F. Kabir, et al. Information theoretic SOP expression minimization technique. Computer and information technology, 2007. iccit 2007. 10th international conference on. IEEE, 2007.

[18] Kabir, Md Faisal, and Simone Ludwig. "Classification of Breast Cancer Risk Factors Using Several Resampling Approaches." 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE, 2018.

[19] Kabir, Md Faisal, and Simone A. Ludwig. "Enhancing the Performance of Classification Using Super Learning." Data-Enabled Discovery and Applications 3.1 (2019): 5.

[20] Loh, WeiYin. "Classification and regression trees." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 1.1 (2011): 14-23.

[21] Fawcett, Tom. An introduction to ROC analysis. Pattern recognition letters 27.8 (2006): 861-874.