Adaptive Feature Selection and Classification of Colon Cancer From Gene Expression Data: an Ensemble Learning Approach

Ashraful Islam
University of Louisiana at Lafayette
Lafayette, Louisiana, USA
ashraful.islam1@louisiana.edu

Mohammad Masudur Rahman
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh
masudurism@gmail.com

Eshtiak Ahmed
Daffodil International University
Dhaka, Bangladesh
eshtiak.cse@diu.edu.bd

Faisal Arafat
Daffodil International University
Dhaka, Bangladesh
faisalarafat.cse@gmail.com

Md Fazle Rabby
University of Louisiana at Lafayette
Lafayette, Louisiana, USA
md-fazle.rabby1@louisiana.edu

ABSTRACT

Cancer research is one of the major and significant areas in medical research. A substantial number of research has been performed in this area and several methods have been employed. However, accuracy of cancer prediction is yet to reach near perfection as the conventional classification methods have several limitations. In recent times, microarray processed gene expression data has been used to predict cancer with significant accuracy. The gene expression data are usually high dimensional and comprises of relatively small number of samples which makes them difficult to classify. In order to achieve higher accuracy, ensembles method can be deployed which combines multiple classification methods. In this study, we have used the public colon cancer gene expression data set that consists of 62 instances having 2,000 attributes. An adaptive pre-processing procedure has been conducted including Linear Discriminant Analysis (LDA) and Principle Component Analysis (PCA) to cope up with the high dimensionality of the data. This was followed by building an ensemble learning model with k-Nearest Neighbors (kNN), Random Forest (RF), Kernel Support Vector Machines (KSVM), eXtreme Gradient Boosting (XGBoost), and Bayes Generalized Linear Model (GLM). Comparing with other classifiers, this study offers a significant improvement as our ensemble learning model gives higher accuracy than previously employed classification techniques. Thus the obtained accuracy is 91.67% with the scores 0.75, 1.00 and 0.85 of precision, recall and Matthews correlation coefficient (MCC) values respectively.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; **Ensemble methods**; • **Applied computing** → **Computational biology**; **Bioinformatics**.

ICCA 2020, January 10–12, 2020, Dhaka, Bangladesh

ACM ISBN 978-1-4503-7778-2/20/01…$15.00
https://doi.org/10.1145/3377049.3377070

KEYWORDS

gene expression, colon cancer, feature selection, DNA micro-array, machine learning, ensemble learning

ACM Reference Format:
Ashraful Islam, Mohammad Masudur Rahman, Eshtiak Ahmed, Faisal Arafat, and Md Fazle Rabby. 2020. Adaptive Feature Selection and Classification of Colon Cancer From Gene Expression Data: an Ensemble Learning Approach. In *International Conference on Computing Advancements (ICCA 2020), January 10–12, 2020, Dhaka, Bangladesh.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3377049.3377070

1 INTRODUCTION

In the field of medicinal science, diagnosis of diseases is a vital and key aspect [1]. Doctors more often than not break down side effects in the human body to foresee illnesses. With a specific end goal to make it more exact, numerous research strategies have been utilized in recent times [2]. Priority has been given to more lethal diseases as they can cause death and widespread investigations have been employed on cancer research. If global statistics of death due to diseases are taken into account, cancer emerges as the second most significant [3]. In the year 2018 alone, cancer has caused an estimated number of 9.6 million deaths [4]. Also, every 1 out of 6 deaths globally is caused by cancer if death by diseases is taken into account [5]. Another survey indicates that every 7 out of 10 deaths due to cancer generally occur in the lower and middle-income regions in the world [3]. To be specific regarding our study, about 1.80 million people are suffering from colorectal (colon) cancer and 47.89% of patients ended with death [4, 6]. This calls for the development of a robust, easy to use and less costly cancer diagnosis method as the mortality rate in cancer can be reduced if they are diagnosed and treated early [3, 4, 7].

In the early stages, cancer classification has dependably been morphological, furthermore, clinical-based [8]. These regular cancer classification techniques are accounted for having a few confinements in their symptomatic capacity [7, 9]. To address the shortcomings of these techniques, researchers started investigating alternative ways [10–12]. It has been proposed that determinations of treatments as indicated by tumor composes separated by pathogenetic examples may boost the viability of the patients [7, 11]. Additionally, the current tumor classes have been observed to be

heterogeneous and involve infections that are molecularly particular and pursue distinctive clinical courses [7, 8, 11, 13]. Advances in the areas of bioinformatics research, particularly in the zone of microarray-based expression analysis have prompted the guarantee of better cancer diagnosis utilizing new molecular-based methodologies; even so, this new sort of data shows new difficulties [6, 9]. On a fundamental level, molecular cancer disease analysis can be scientifically detailed as classification errands. Most past examinations around there finished algorithmic answers for binary classification, e.g. tumor versus typical tissue, positive treatment versus no reaction [6–9, 14]. In light of the vast number of tumor composes what's more, subtypes, it is basic need to create multi-class tumor distinguishing techniques for practical malignancy determination purposes [7, 8, 10]. The ongoing advent of DNA microarray innovation permits the genomic-scale investigation of natural procedures through synchronous checking of the relative expression values of thousands of qualities. With this plenitude of gene expression data, scientists have begun to investigate the conceivable outcomes of cancer characterization. A significant number of techniques have been proposed as of late with promising outcomes [2, 7, 12, 14–16].

Conventional cancer analysis depends on an intricate and estimated blend of clinical and histo-pathological information. These classic methodologies may come up short when managing atypical tumors or morphologically vague tumor sub-types. Likewise, it is challenging to utilize quality gene expression data for cancer classification due to the accompanying two special parts of quality gene expression data [15, 17]. Initially, quality gene expression data are generally high dimensional having a range between a thousand and ten thousand, sometimes even more. Second, quality gene expression data sets normally contain a small number of quantities of test samples, sometimes less than 70. So it is a challenge to handle those data properly to estimate an accurate prediction [14, 16, 18]. Generally, microarray-based gene expression data classification task requires two stages i.e. feature selection task followed by the data classification task. As the gene expression datasets usually comprise more than a few thousand genes, it is crucial to choose the genes in the feature selection process [16]. While picking the features from the data, the advantages of some traditional dimensionality reduction methods i.e. principal component analysis (PCA), linear discriminant analysis (LDA) have been performed [15]. Alongside the feature selection techniques, some robust methods i.e. artificial neural network (ANN), support vector machine (SVM), k-nearest neighbor (kNN) have been utilized comprehensively in microarray-based gene expression data classification [18]. Later, incorporating several traditional classification algorithms for building a committee referred as ensemble learning methods have offered better accuracy along with greater reliability.

In the beginning period of this kind of research, the model proposed by Backert et al. could achieve an accuracy of nearly 50% marginally over a dataset containing 588 gene expressions acquired from colon tissues [19]. However, a multiple kernel support vector machine (MK-SVM) classifier achieved more than 90% accuracy which was employed by Chen and Li on colon tumor and leukemia data sets [20]. Likewise, Li et al. utilized a genetic algorithm (GA) to distinguish discriminative genes and kNN for classifying the data which led to an accuracy of 94.1% [21]. Research proceeded in this space, and the accuracy of 92% acquired by characterizing colon cancer data using a probabilistic neural network (PNN) by Shon et al. was promising for further investigations [22]. Furthermore, another method for identifying colon cancer was proposed by Kulkarni et al. and their model combined decision trees (DTs) and genetic programming for classification task which showed an outcome of greater extent in performance over the previously available techniques [12]. As of late, the proposed strategy made by Tong et al. had the most conceivable performance in colon cancer recognition due to their capability of choosing an optimal assembly of linear SVM classifiers for a data set of 50 gene expressions [23].

While different classifiers have different types of advantages, combining some of them to make use of all the best features is a very potent possibility. This study aims to construct an ensemble classifier which can classify the normal tissues and abnormal tissues (cancerous tissues) more reliably from the microarray-based gene expression profile available in the public colon cancer gene expression data set. This data set consists of 62 instances having 2,000 attributes has been collected from the Bioinformatics Research Group of Escuela PolitÃĺcnica Superior, Spain. Further preprocessing works have been conducted by applying LDA, and PCA followed by building an ensemble learning model with the typical machine learning (ML) algorithms e.g. kNN, Random Forest (RF), Kernel Support Vector Machines (KSVM), eXtreme Gradient Boosting (XGBoost), Bayes Generalized Linear Model (GLM). This ensemble learning model gives more accuracy than other classifiers found in the literature review which is 91.67% with the scores 0.75, 1.00 and 0.85 of precision, recall, and Matthews correlation coefficient (MCC) values respectively.

The layout of this paper is organized as follows- a brief description of necessary preliminaries is presented in section 2, followed by an overview of the data set in section 3. Then the steps required for conducting this study are described in section 4. Section 5 explains the results and discussions established in this study. Finally, the article is concluded along with some implications for further study in this area in section 6.

2 PRELIMINARIES

For understanding this works properly, this section provides potential supporting knowledge and materials about microarray and gene expression data along with a brief description on fundamentals of ensemble learning.

2.1 Microarray and Gene Expression Data

DNA microarray can be employed for forming gene expression data [10, 14, 24]. Generally DNA microarray points to a collection of different spots on a glass slide where every spot may have a few millions of DNA molecules those are identical. This identical DNA molecules fit to specific genes and provide an image as an output in which every gene has a corresponding location. This locations represent the relative expression level based on the associated fluorescenes values for the respective genes. Then the gene expression values are being measured to store in a matrix where the genes and corresponding expression values based on various conditions are represented by the rows and the columns respectively [17]. Figure 1 shows a symbolic matrix representation of gene expression data

in which the number of rows (genes) is n whilst the number of columns (expression values based on conditions) is m.

2.2 Ensemble Learning

ML is a branch of artificial intelligence which is employed for learning from given data (training data) over the given conditions and then it can conclude with an inferred decision for test data [1, 7, 25, 26]. Generally ML is classified into two types which are known as-

- **Supervised Learning:** A labeled training data set is being used to estimate a decision over the test data [26].
- **Un-supervised Learning:** A decision is being estimated based on an unlabeled training data set [26].

The main task of ML is to build a model which can be trained using existing available data so that it can estimate a new decision over a specific region of rules and scopes. An ML model usually evaluated based on its performance measures i.e. accuracy of prediction [1, 25, 26] . But sometimes a committee of several ML algorithms (classifiers) can perform better than employing a single algorithm (e.g. KNN) and this committee can be formed by several different algorithms according to the target goal. Technically this committee of such algorithms is known as ensemble learning method and task main task of building this committee is to boost the overall performances [26]. A conceptual view of a generalized ensemble learning method is available in Figure 2.

3 DATA SET

This section provides an overall view of the data set which has been experimented on in this study. The data set used in this research is collected from the Bioinformatics Research Group of Pablo de Olavide University, Spain [27]. These sample tissues have been collected from the biopsies of various patients who have been suffering from colon cancer. Initially, the number of collected genes are 6,817 which derives 7,129 probes. These raw samples are clinically processed to generate gene expression data. After processing, 62 samples are collected with 2,000 attributes where the attributes consist different values of related genes responsible for determining whether a tissue is cancerous or not. Among the samples, 40 instances were captured from tumor cells which are labelled as *positive* while other 22 instances are labelled as *negative* as they remained normal cells in the same patients. The attributes are the features in this study which are being employed to train the model and then classify the test data.

4 METHODS

This section provides an extensive description of methods used to perform the experiment sequentially according to the steps. An overview of the whole process performed in this study can be perceived through the flow diagram illustrated in Figure 3 before driving to the description of related steps mentioned as follows-

4.1 Step 1: Pre-processing

As the gene expression data is used for this study, some sorts of modular analysis are required to analyze and process the data before they are undergoing ML methods [25, 28, 29]. The pre-processing task starts with a statistical analysis of the collected data. The central tendency measurements, different kinds of histograms, and plots help and guide to extract useful information and patterns which contains valid unique identifiers and conditions [12, 16, 28]. Then the inconsistent replicated gene IDs are spotted and deducted from the data set. On the following, the data are filtered to find out the flat patterns by calculating the root mean square (RMS) and standard deviation of the gene expression patterns. Another issue with the data set is the fact that there were missing values. To solve this, gene expression patterns with too many unknown values are deleted intuitively and imputed in case of the remaining less frequent missing ones. The central tendency measures is utilized while imputing using Eq. 3 and the quality of imputed data is maintained by calculating *Relative Absolute Error (RAE)* using Eq. 1 [30]. Since the ratios of statistical analysis are symmetrical, the data set are log-transformed to normalize the values after imputing missing values applying Eq. 3.

$$RAE = \frac{1}{number\ of\ missing\ data} \sum_{y_{gs}\ missing} \frac{|\hat{y}_{gs} - y_{gs}|}{\phi(y_{gs})} \quad (1)$$

where,

$$\phi(y_{gs}) = \begin{cases} |y_{gs}| & if\ \ |y_{gs}| > \epsilon \\ \epsilon & if\ \ |y_{gs}| < \epsilon \end{cases} \quad (2)$$

$$\hat{y}_{gs} = \frac{\sum_{i=1}^{n} r_{is} x_{is}}{\sum_{i=1}^{n} r_{is}} \quad (3)$$

where, x_{is} is the expression value and $r_{is} = 1$ if x_{is} is observed while $r_{is} = 0$ if x_{is} is missing.

4.2 Step 2: Adaptive Feature Selection

Feature selection methods can be of two types: filtering method which ranks the features based on the contribution ignoring feature dependencies, and wrapper method which splits the features into subset to make them optimal for good classification [12, 16, 31]. A composition of both approaches in an adaptive way have been used which is a novelty of our research. Before experimenting the proposed adaptive feature selection process, basic PCA and LDA on our data set have been applied to shrink features on the basis of dimension reduction principle [31].

Our adaptive feature selection starts with whole data set (2,000 features) and a regular classifier like DT. On the creation of RF tree, filtering feature selection has been employed for measuring a contribution score in percentile. Also, the features with less contributing score are discarded considering a threshold of 5%. After the modeling, less features had remained to move on to building the next classifiers. Subsequently, four more classifiers have been experimented in the same way with selected features. Table 1 summarizes number of selected features for the classifiers individually.

4.3 Step 3: Classification

With the selected features, the classificartion process was initiated with engaging single ML methods individually and ensemble methods later, on the following, to build the model. Five typical ML algorithms were employed for building ensemble method (model).

Figure 1: A symbolic $n \times m$ matrix representation of a gene expression data.

Template/acmart-master/images/gene$_e$*xpression.png*

		Conditions				
		Condition-1	*Condition-2*	...	*Condition-(m-1)*	*Condition-m*
Gene_id	*Gene-1*	301.51	876.45	...	12054.87	7691.38
	Gene-2	1423.13	643.96	...	1532.54	983.11
	...	...	...	...	...	...
	Gene-(n-1)	5000.98	983.11	...	32.91	668.62
	Gene-n	93.09	1532.54	...	876.45	643.96

Figure 2: A typical process of building an ensemble model for boosting performance.

Template/acmart-master/images/ensemble.jpg

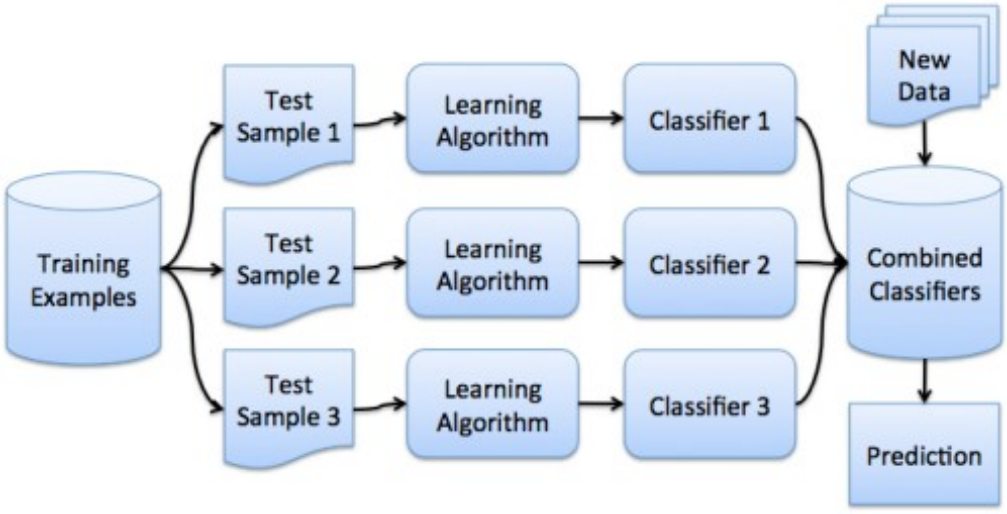


Table 1: Summary of proposed adaptive feature selection.

Classifier	Selection Method	Number of Selected Features
DT	Filtering	1,800
SVM	Wrapper	1,750
kNN	Filtering	1,600
RF	Filtering	1,200

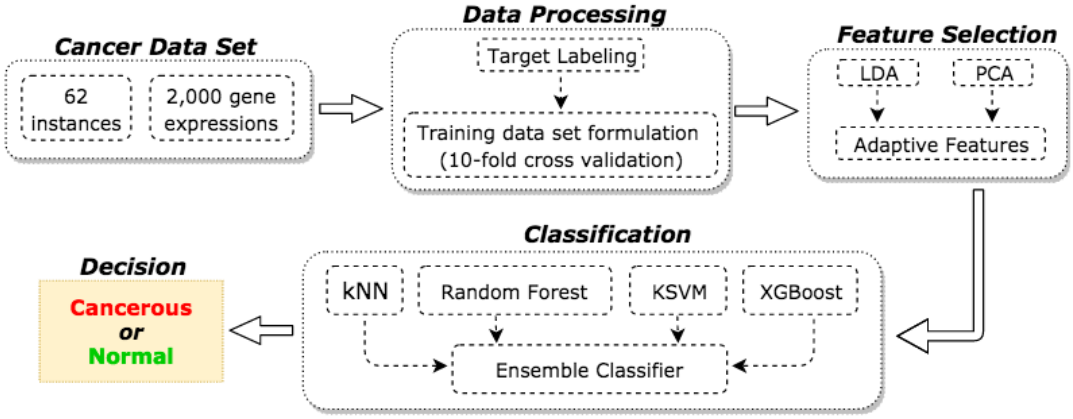
This algorithms have been chosen based on their individual performances and objectives relative to current data set. The proposed ensemble model comprises of the following ML algorithms: kNN, RF, KSVM, XGBoost, and GLM.

4.4 Step 4: Performance Evaluation

Unlike other ML approaches, the results of our experiments are assessed using well-known performance measures such as accuracy, precision, recall, and MCC. Additionally, the area under curve (AUC) in receiver operating characteristic (ROC) have been visualized for recognizing the best classifier involve in this stage. A 10-fold cross validation on the 80% data and the rest 20% data is used for testing. As the problem in hand is a binary classification problem, TP, TN, FP and FN terms were used which denote the number of correctly classified positive instances of data set, the number of correctly classified negative instances, the number of incorrectly classified positive instances and the number of incorrectly classified negative instances respectively [26].

Figure 3: Overall flow diagram of related steps and methods used in this study.

Template/acmart-master/images/Flowchartv6.png



4.4.1 Accuracy: It is a basic performance metric which assesses the overall *effectiveness* of a classifier and calculated using Eq. 4

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{4}$$

4.4.2 Precision: It is a measure of *exactness* which states precisely the positive predictions using the formula in Eq. 5

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

4.4.3 Recall: It refers the true positive rate which finds the proportion of correct positive classification those are actually positive. It is determined by Eq. 6

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

4.4.4 MCC:. It is an efficient metric for evaluating binary classifier unlike our ones. It is extremely important when the classes are imbalanced. The range of MCC value lies in between -1 and +1 where -1 indicates extreme mismatch between actual and predicted class and +1 denotes perfect prediction. It if formulated by using Eq. 7.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{7}$$

5 RESULTS AND DISCUSSION

The following Table 2 shows the comparison of different singleton methods with our proposed ensemble technique. It is observed that our ensemble method yields higher accuracy of 91.67% than other individual methods. It is noted that the results are obtained from 10-fold cross validation of the data set. DT, Naive Bayes (NB), RF and AdaBoost methods have been implemented individually for measuring accuracy. In the end, these accuracy numbers were considered while building the ensemble classifier which consists of kNN, RF, KSVM, XGBoost, and GLM. Figure 4 represents the ROC with AUC at 95% confidence interval of the best performer classifier in the ensemble method.

Table 2: Performance Evaluation Results

Metric	*DT*	*NB*	*RF*	*AdaBoost*	*Ensemble*
Accuracy	80.6	74.2	85.5	88.7	91.67
Precision	73.3	60.7	76	84.6	82
Recall	84.6	77.3	91.7	88	100
MCC	0.616	0.48	0.715	0.767	0.85

Our experiments have been conducted using R programming language on a desktop computer with 2.4 GHz Intel Core i7 processor, 8 GB RAM and Ubuntu 16.04 64-bit operating system installed. Some core ML and ensemble leaning algorithms were used in our experiments. The coding implementation of ML algorithms are available in different R packages namely- *e1071, rpart, randomForest,* and *caret.* The parameters of the default code have been changed in the interest of enhancing performance of the models.

On the purpose of ensemble learning method, [32] package of R language has been used which facilitates the flexibility of using different ensemble techniques with amalgamation. The parameters of *SuperLearner* are altered and in some cases several codes are overridden for the sake of experiment.

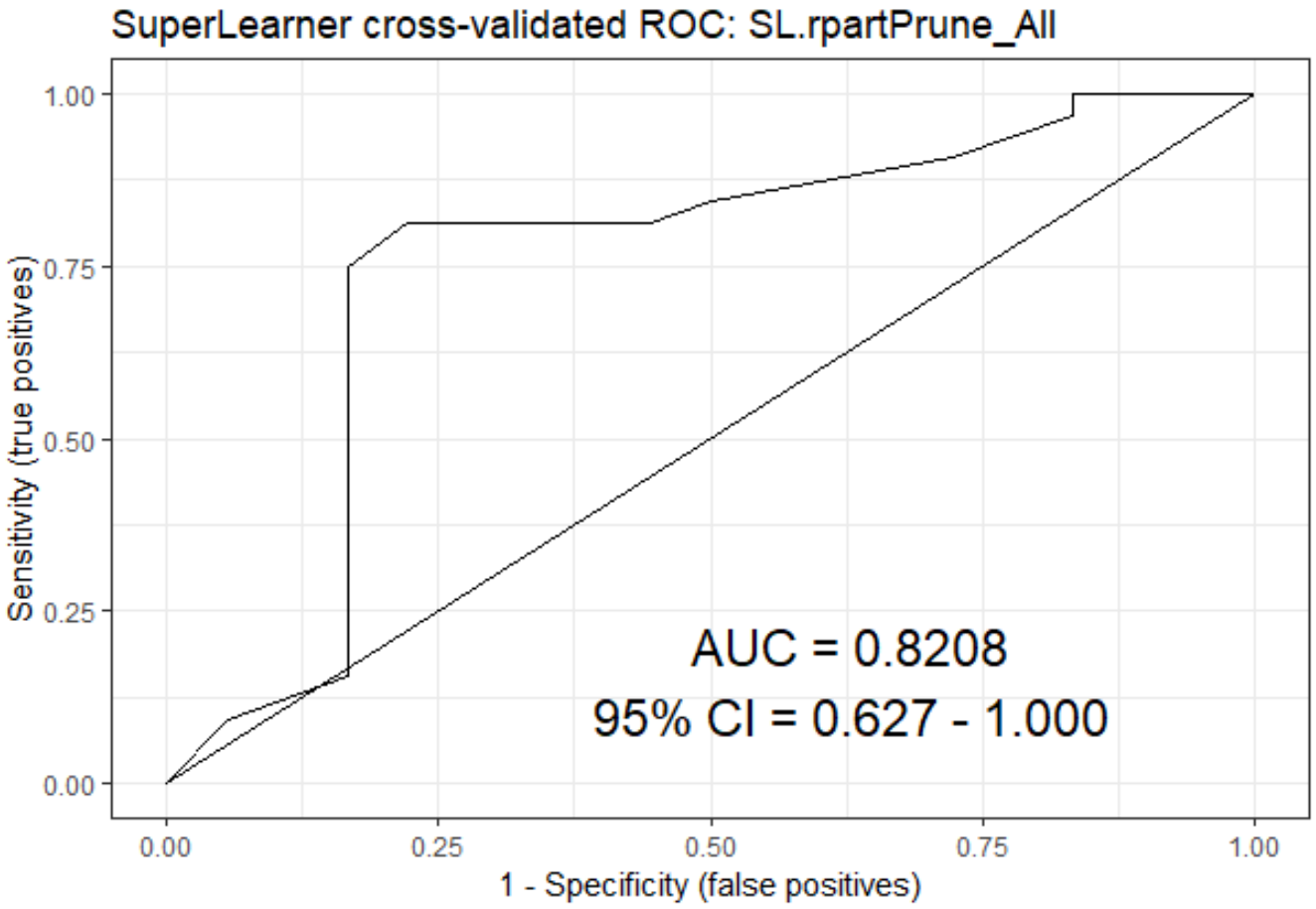
To implement our proposed feature selection technique, the whole feature vector has been examined with several algorithms to find the important features. As discussed earlier, the number of feature are reduced from 2,000 to 1,200, a total of almost 50 models with different parameters are performed and finally the purpose is served.

While evaluating the performance measures- accuracy, precision, and recall are default functions in *dplyr* package which helps to build the confusion matrices while MCC was externally programmed in R. Also, *ROCR* and *ck37r* packages have been used to generate ROC curve and other necessary plots of the experiment.

6 CONCLUSIONS

Over the years in cancer prediction and classification researches, it has been experienced that 100% accuracy cannot be promised

Figure 4: ROC curve of XGBoost algorihm which performs best in the ensemble method.
Template/acmart-master/images/ROC.png



as human-intervention is much required from the medical practitioners and experts. Although respective to this experiment, the accuracy is higher than previous results, it is limited to a certain number of instances in data set used which is quite small for being trustworthy. But this study has shown a significant hope that further investigations with larger data set could have a great contribution both in theoretical and practical aspects in colon cancer prediction and classification. Meanwhile, deep-learning based ML classifiers would be a potential choice for greater accuracy on the other hand. Therefore, this higher accuracy shows a sign that investigating ML methods on gene expression data are potent in cancer classification, specifically in colon cancer.

REFERENCES

[1] Min Chen, Yixue Hao, Kai Hwang, Lu Wang, and Lin Wang. Disease prediction by machine learning over big data from healthcare communities. *IEEE Access*, 5: 8869–8879, 2017.

[2] Kun Lan, Dan-tong Wang, Simon Fong, Lian-sheng Liu, Kelvin KL Wong, and Nilanjan Dey. A survey of data mining and deep learning in bioinformatics. *Journal of medical systems*, 42(8):139, 2018.

[3] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L. Siegel, Lindsey Torre, and Ahmedin Jemal. Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *Global Cancer Statistics 2018*, 2018.

[4] World Health Organization. Global Health Observatory (GHO): Data Repository, 2018. URL http://apps.who.int/gho/data/node.home.

[5] American Cancer Society. Cancer Facts and Figures 2017. *Genes and Development*, 21(20):2525–2538, 2017. ISSN 08909369. doi: 10.1101/gad.1593107.

[6] Elizabeth Alwers, Min Jia, Matthias Kloor, Hendrik Bläker, Hermann Brenner, and Michael Hoffmeister. Associations between molecular classifications of colorectal cancer and patient survival: A systematic review. *Clinical Gastroenterology and Hepatology*, 2018.

[7] Konstantina Kourou, Themis P Exarchos, Konstantinos P Exarchos, Michalis V Karamouzis, and Dimitrios I Fotiadis. Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, 13: 8–17, 2015.

[8] Chen-Hsiang Yeang, Sridhar Ramaswamy, Pablo Tamayo, Sayan Mukherjee, Ryan M Rifkin, Michael Angelo, Michael Reich, Eric Lander, Jill Mesirov, and Todd Golub. Molecular classification of multiple tumor types. *Bioinformatics*, 17 (suppl_1):S316–S322, 2001.

[9] Rick Kamps, Rita D Brandão, Bianca J Bosch, Aimee DC Paulussen, Sofia Xanthoulea, Marinus J Blok, and Andrea Romano. Next-generation sequencing in oncology: genetic diagnosis, risk prediction and cancer classification. *International journal of molecular sciences*, 18(2):308, 2017.

[10] S Rathore, M Hussain, and A Khan. GECC: Gene Expression Based Ensemble Classification of Colon Samples. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 11(6):1131–1145, 2014. ISSN 15455963.

[11] Y. Liu, J. Zhou, and Y. Chen. Ensemble classification for cancer data. In *BioMedical Engineering and Informatics: New Development and the Future - Proceedings of the 1st International Conference on BioMedical Engineering and Informatics, BMEI 2008*, volume 1, 2008. ISBN 9780769531182. doi: 10.1109/BMEI.2008.161.

[12] Ashwinikumar Kulkarni, B. S C Naveen Kumar, Vadlamani Ravi, and Upadhyayula Suryanarayana Murthy. Colon cancer prediction with genetics profiles using evolutionary techniques, 2011. ISSN 09574174.

[13] S Backert, M Gelos, U Kobalz, M L Hanski, C Böhm, B Mann, N Lövin, A Gratchev, U Mansmann, M P Moyer, E O Riecken, and C Hanski. Differential gene expression in colon carcinoma cells and tissues detected with a cDNA array. *International journal of cancer*, 82(6):868–74, 1999. ISSN 0020-7136.

[14] Xuan Long, Zhigang Deng, Guoqiang Li, and Ziwei Wang. Identification of critical genes to predict recurrence and death in colon cancer: integrating gene expression and bioinformatics analysis. *Cancer cell international*, 18(1):139, 2018.

[15] Hala Alshamlan, Ghada Badr, and Yousef Alohali. mrmr-abc: a hybrid gene selection algorithm for cancer classification using microarray gene expression profiling. *BioMed research international*, 2015, 2015.

[16] Zena M Hira and Duncan F Gillies. A review of feature selection and feature extraction methods applied on microarray data. *Advances in Bioinformatics*, 2015, 2015.

[17] Uri Alon, Naama Barkai, Daniel A Notterman, Kurt Gish, Suzanne Ybarra, Daniel Mack, and Arnold J Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, 96(12):6745–6750, 1999.

[18] Huaming Chen, Hong Zhao, Jun Shen, Rui Zhou, and Qingguo Zhou. Supervised machine learning model for high dimensional gene data in colon cancer detection. In *Big Data (BigData Congress), 2015 IEEE International Congress on*, pages 134–141. IEEE, 2015.

[19] Steffen Backert, Marcos Gelos, Ursula Kobalz, Marie-Luise Hanski, Christian Böhm, Benno Mann, Nicole Lövin, Alexei Gratchev, Ulrich Mansmann, Mary Pat Moyer, et al. Differential gene expression in colon carcinoma cells and tissues detected with a cdna array. *International journal of cancer*, 82(6):868–874, 1999.

[20] Zhenyu Chen and Jianping Li. A multiple kernel support vector machine scheme for simultaneous feature selection and rule-based classification. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 441–448. Springer, 2007.

[21] Leping Li, Clarice R Weinberg, Thomas A Darden, and Lee G Pedersen. Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the ga-knn method. *Bioinformatics*, 17(12): 1131–1142, 2001.

[22] Ho Sun Shon, Gyoyong Sohn, Kwang Jung, Sang Yeob Kim, Eun Jong Cha, and Keun Ryu. A gene expression data classification using discrete wavelet transform. In *International Conference on Bioinformatics & Computational Biology (BIOCOMP2009)*, pages 204–208, 2009.

[23] Muchenxuan Tong, Kun Hong Liu, Chungui Xu, and Wenbin Ju. An ensemble of SVM classifiers based on gene pairs. *Computers in Biology and Medicine*, 43(6): 729–737, 2013. ISSN 00104825.

[24] Aik Choon Tan and David Gilbert. Ensemble machine learning on gene expression data for cancer classification. *Applied Bioinformatics*, 2(3 Suppl):S75–83, 2003. ISSN 1175-5636.

[25] Jasmina Dj Novakovic, Alempije Veljovic, Sinisa Ilic, Zeljko Papic, and Milica Tomovic. Evaluation of Classification Models in Machine Learning. *Theory and Applications of Mathematics & Computer Science*, 7(1):39–46, 2017.

[26] Micheline Kamber Jiawei Han and Jian Pei. *Data Mining: Concepts and Techniques*. The Morgan Kaufmann, 2007.

[27] Spain Bioinformatics Research Group, Pablo de Olavide University. Colon cancer dataset in arff format with labelled classes., 2018. http://eps.upo.es/bigs/datasets.html.

[28] R. Diaz-Uriarte J. Herrero and J. Dopazo. Gene expression data preprocessing. *Bioinformatics Applications Note*, 19(5):655âĂŞ656, 2003.

[29] F Alexander Wolf, Philipp Angerer, and Fabian J Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19(1):15, 2018.

[30] Naisyin Wang Danh V. Nguyen and Raymond J. Carroll. Evaluation of missing value estimation for microarray data. *Bioinformatics Applications Note*, 19(5): 347–370, 2004.

[31] Yvan Saeys, IÃśaki Inza, and Pedro Larranaga. A review of feature selection techniques in bioinformatics. *Bioinformatics (Oxford, England)*, 23:2507–17, 11 2007.

[32] Eric Polley. Superlearner package. https://cran.r-project.org/web/packages/SuperLearner/SuperLearner.pdf., 2 September, 2018.