PERFORMANCE ANALYSIS OF FEATURE RANKING ALGORITHMS ON MICROARRAY DATASETS

Uğur Turhal[1], Murat Gök[2], Suat Onur[3], Sebahattin Babur[4]

[1,2,4]Department of Computer Engineering
[3]Department of Informatics,
[1,3] BalıkesirUniversity
[2,4] Yalova University

[1] *ugurturhal@balikesir.edu.tr* [2] *murat.gok@yalova.edu.tr*
[3] *suatonur@balikesir.edu.tr* [4] *sebahattin_babur@hotmail.com*

ABSTRACT

The microarray datasets host a lot of information which influence the problems with different the degree. Choosing the minimum number of features (attributes) which are representing of these data structures as an optimization problem. Nowadays, the microarray datasets are utilized in the diagnose of cancer diseases. However, their size may cause the curse of dimensionality for machine learning methods during classification(Loris, N. et al., 2012). Therefore, they need more computing power and long processing times. Hence, reducing the number of attributes will be fundamental step to solve this problem. In this study, "Colon" and "Ovarian" datasets which are used frequently in literature were processed with various feature ranking algorithms. The best “k” number features, which chosen after ranking were classified with "Naive Bayes” and "SVM(Linear) classifiers. The evaluation of the system was realized on "Kappa", "MCC" and "Accuracy" scores and "ROC" graphs. This study aims to provide helpful information to the researchers who work on the same datasets.

Keywords: Microarray datasets, Feature ranking, Naive Bayes, SVM

I. INTRODUCTION

DNA microarray technology has proven to be an important breakthrough in molecular biology. This rapidly maturing technology is providing scientists with a means of monitoring the expression of genes on a genomic scale(Chee, M.*et al.* 1996).

Cancer is a broad group of diseases involving unregulated cell growth. In cancer, cells divide and grow uncontrollably, forming malignant tumors, which may invade nearby parts of the body. Not all tumors are cancerous; benign tumors do not invade neighboring tissues and do not spread throughout the body. There are over 200 different known cancers that affect humans (Cancer Research UK, 2012).

In 2007, cancer caused about 13% of all human deaths worldwide (7.9 million). Rates are rising as more people live to an old age and as mass lifestyle changes occur in the developing world (Jemal A, *et al.* 2011). According to American Cancer Society, about 1,665,540 new cancer cases are expected to be diagnosed and about 585,720 of them are expected to die in America, 2014(American Cancer Society, 2014).

The American men-women who died owing to different cancer diseases between 1930 and 2010 are shown in the following figures I-II.

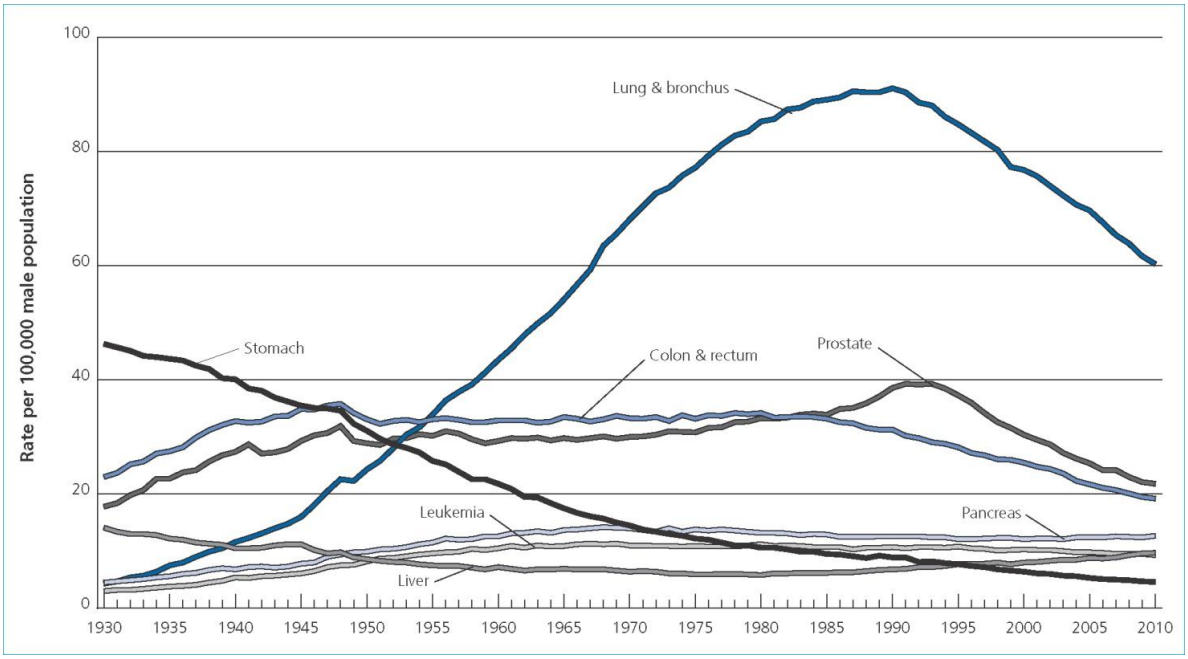


***Figure I**: Age-adjusted Cancer Death Rates, Males by Site, US, 1930-2010*(American Cancer Society, 2014).

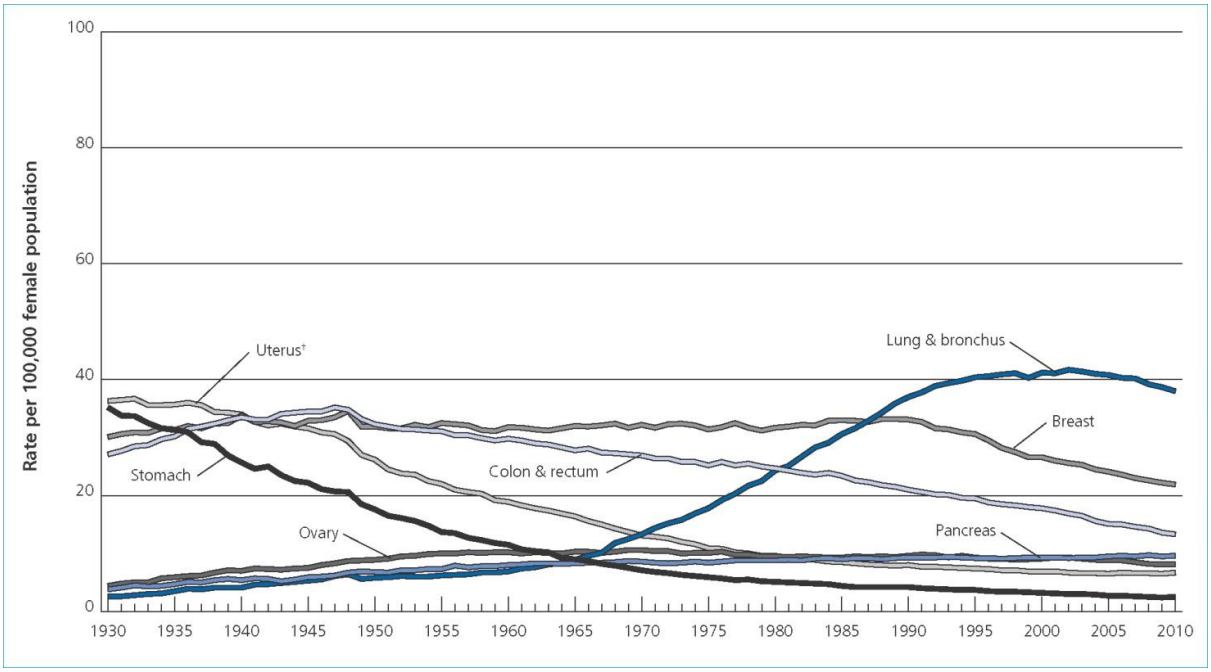


***Figure II**: Age-adjusted Cancer Death Rates, Females by Site, US, 1930-2010*(American Cancer Society, 2014).

The microarray data sets host a lot of information which influence the problems with different the degree. One of important application area is disease prognostication(Golub, T.R. *et al.* 1999).Hence, choosing the minimum number of features (attributes) which are representing of these data structures as an optimization problem.

In our former studies, we have improved the performance of classification with using ensemble classification methods on "Colon" and "Thyroid" microarray datasets(Akbaş, A. *et al.* 2013;Babur, S. *et al.* 2012;Turhal, U. *et al*. 2013). In this study, "Ovarian" and "Colon"datasets which are used frequently in literature were processed with various feature ranking algorithms. The best "k" (150 and 300) number features, which chosen after ranking were classified with "Naive Bayes" and "SVM(Linear)" classifiers. The evaluation of the system was realized on "Kappa", "MCC" and "Accuracy" scores and "ROC" graphs.

Finally all results have been compared and best ranking methods and classifiers for each datasets are shown in the tables.

II. MATERIAL AND METHODS

In this study, several experiments have been conducted on 2 publicly available datasets. Below were provided a brief description for each dataset. (the salient features of each dataset are summarized in ***Table I***):

***Table I:** Characteristics of the datasets used in the experiments: the first column presents the number of features (#F), and the second column reports the number of samples (#S)*(Loris, N. *et al.*2012).

Dataset	**#F**	**#S**
Ovarian (O)	15154	253
Colon (C)	2000	62

Ovarian dataset (O): the ovarian dataset contains 253 samples and two class are considered: 91 samplesare normal and 162 samplesare ovarian cancers(Petricoin,E.F. *et al.* 2002);

Colon (C): the colon dataset contains 62 samples and two class are considered: 22 samples are normal and 40samples are tumor cancers(Alon,U. *et al.*1999);

A. Feature Ranking

Many feature ranking methods are using frequently in literature. However all methods have advantages and disadvantages while comparing each others. All feature ranking methods that used in this study are described below;

1. *Bhattacharyya*

The Bhattacharyya coefficient is an approximate measurement of the amount of overlap between two statistical samples. The coefficient can be used to determine the relative closeness of the two samples being considered. It is calculated by following equation(Djouadi, A. *et al. 1990);*

$$Bhattacharyya = \sum_{i=1}^{n} \sqrt{(\sum a_i \; x \; \sum b_i)} \quad (1)$$

Where,

a, b : samples

n : number of partitions

$\sum a_i$, $\sum b_i$: numbers of members of samples a and b in the i_{th} partition.

2. *T-Test*

T-test is one method for testing the degree of difference between two means in small sample. It uses T distribution theory to deduce the probability when difference happens, then judge whether the difference between two means is significant (Jiaxi, L. 2010). It is calculated by following equation;

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \quad (2)$$

Where,

$\overline{x_1}$ = Average of first set of values

S_1 = Standard deviation of first set of values

n_2 = Total number of values in first set

$\overline{x_2}$ = Average of second set of values

S_2 = Standard deviation of second set of values

n_2 = Total number of values in second set

3. *Wilcoxon*

Absolute value of the standardized u-statistic of a two-sample unpaired Wilcoxon test, also known as Mann-Whitney U test, is a non-parametric test of the null hypothesis that two populations are the same against an alternative hypothesis, especially that a particular population tends to have larger values than the other (Wilcoxon, F. 1945).It is calculatedwith two formulas below (Mann, H.B. and Whitney, D.R. 1947);

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2} \quad (3)$$

$$U_2 = R_2 - \frac{n_2(n_2+1)}{2} \quad (4)$$

Where,

n_1: the sample size for sample 1
R_1: the sum of the ranks in sample 1
U_1: observation and the total ranking number for sample 1
n_2: the sample size for sample 2
R_2: the sum of the ranks in sample 2
U_2: observation and the total ranking number for sample 2

B. Feature Selection

In this section, the features of microarray datasets that used in the work are ranked according to significance level. After that, first k number features are selected and created a new dataset. Feature selection process is repeated for k=150 and k=300.

C. Classifiers

The classifiers used in this study are described below;

1. *Naïve Bayes*

Naive Bayes is the simplest form of Bayes Net. All features are independent from given class variables. This method is called conditional independency (Zhang, H. 2005).

$$f_{nb}(E) = \frac{p(C=+)}{p(C=-)} \prod_{i=1}^{n} \frac{p(x_i|C=+)}{p(x_i|C=-)} \quad (5)$$

2. *Support Vector Machines (with Linear Kernel)*

The *support vector machine* or SVM, first described by Vapnik and collaborators in 1992(Boser, B.E. *et al.* 1992), has rapidly established itself as a powerful algorithmic approach to the problem of classification within the larger context known as supervised learning (William H. 2007).

D. Performance Measurement

In order to increase reliability of results, some evaluation methods have been used that found acceptance in literature. These methods;

1. *Accuracy (Acc)*

The accuracy of a measurement system is the degree of closeness of measurements of a quantity to that quantity's actual (true) value (Taylor, R. 1999). It is calculated by following equality;

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (6)$$

Where,

TP : Number of real positives
TN : Number of real negatives
FP : Number of unreal positives
FN : Number of unreal negatives

2. Kappa

Cohen's kappa coefficient is a statistical measure of inter-rater agreement or inter-annotator agreement for qualitative items (Cohen, J. 1960). Bigger difference means better result. It is calculated by following equality;

$$K = \frac{\Pr(a)-\Pr(e)}{1-\Pr(e)} \quad (7)$$

$\Pr(a)$: Adding proportion of observed compatibilities for two data,

$\Pr(e)$: Probability of emergence by coincidence for this compatibility

K : Kappa result

3. Matthews Correlation Coefficient (MCC)

The measure was introduced in 1975 by Matthews (Matthews, B.W. 1975).The Matthews correlation coefficient (MCC) is using as a measure of the quality of binary (two-class) classifications.Bigger difference means better result. It is calculated by following equation;

$$MCC = \frac{TPxTN-FPxFN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (8)$$

$TP, TN, FP \text{ and } FN$ are explained under the Accuracy header.

4. ROC

It is a method used for showing performance of binary classifier with graphic (Swets, A. 1996). It is calculated by following equation;

$$ROC = \frac{sensitivity}{1-specifity} \quad (9)$$

Where,

$$Sensitivity : TPR = \frac{TPR}{P} = \frac{TP}{(FP+FN)} \quad (10)$$

$$Specifity : SPC = \frac{TN}{N} = \frac{TN}{(FP+TN)} \quad (11)$$

$TP, TN, FP \text{ and } FN$ are explained under the Accuracy header.

E. Classification and Results

The datasets that obtained in section **B** are classified with classifiers which described in section **C**. Ten-fold cross-validation method was used during the classification. The obtained outcomes are shown in the tables.

The accuracy results that obtained by the raw datasets are shown in the ***Table II***.

Table II: *The accuracy results of full datasets.(%)*

	Ovarian k = 15154	**Colon k = 2000**
Naive Bayes	92,4901	53,2258
SVM (Linear)	100,0000	82,2581

This results show that Linear SVM is better than the Naive Bayes for each dataset. This is because the Linear SVM is appropriate to the large size datasets (McCue, R. 2009). Classification performance results of the best 150 features for each datasets are shown the tables below. The most effective values are shown bold in a yellow cell.

Table III: *Ovarian dataset results (feature count "k" = 150)*

Ovarian	**NaiveBayes**			**SVM - Linear**		
k = 150	**Acc (%)**	**MCC**	**Kappa**	**Acc (%)**	**MCC**	**Kappa**
bhattacharyya	**98,4190**	0,966	0,9655	**100,000**	1,000	1,0000
ttest	97,6285	0,949	0,9480	**100,000**	1,000	1,0000
wilcoxon	88,5375	0,761	0,7576	99,2095	0,983	0,9829

Table IV: *Colon dataset results (feature count "k" = 150)*

Colon	**NaiveBayes**			**SVM - Linear**		
k = 150	**Acc (%)**	**MCC**	**Kappa**	**Acc (%)**	**MCC**	**Kappa**
bhattacharyya	**82,2581**	0,656	0,6384	79,0323	0,547	0,5467
ttest	75,8065	0,560	0,5250	**80,6452**	0,587	0,5857
wilcoxon	72,5806	0,453	0,4411	69,3548	0,352	0,3506

May be reached the following outcomes by referencing the above values;

✓ In all datasets, the highest results for Naive Bayes classifier were obtained by using bhattacharyya method.

✓ In Ovarian dataset, the highest results of best 150 features were obtained by using Linear SVM classifier.

The ROC graphs of the above classification results are given below;

Figure III: *Ovarian dataset ROC graph (feature count "k" = 150)*

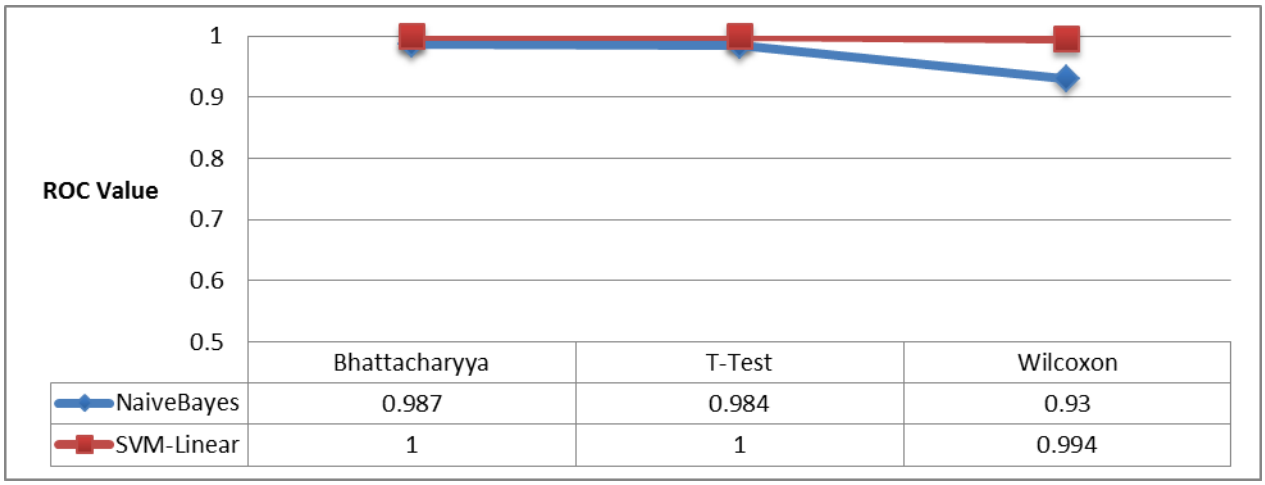
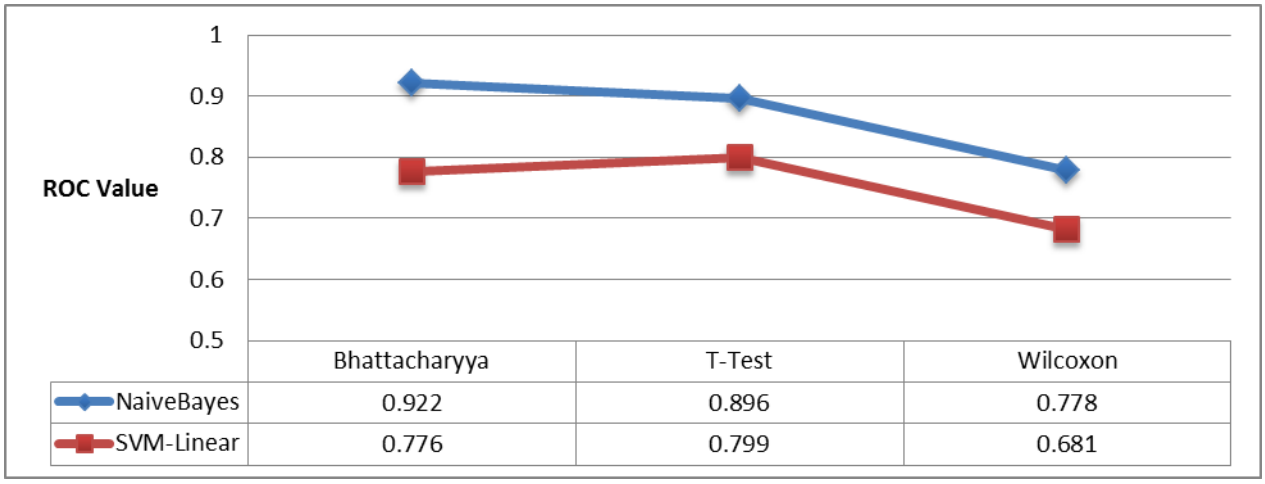


Figure IV: *Colon dataset ROC graph (feature count "k" = 150)*



The classification results and ROC graphs of first 150 feature are given above. The results of the best 300 features are given below.

Table V: *Ovarian dataset results (feature count "k" = 300)*

Ovarian	**NaiveBayes**			**SVM - Linear**		
k = 300	**Acc (%)**	**MCC**	**Kappa**	**Acc (%)**	**MCC**	**Kappa**
bhattacharyya	96,4427	0,923	0,9226	**100,0000**	1,000	1,0000
ttest	**96,8379**	0,931	0,9310	**100,0000**	1,000	1,0000
wilcoxon	83,3992	0,656	0,6514	97,2332	0,941	0,9404

Table VI: *Colon dataset results (feature count "k" = 300)*

Colon **k = 300**	**NaiveBayes**			**SVM - Linear**		
	Acc (%)	**MCC**	**Kappa**	**Acc (%)**	**MCC**	**Kappa**
bhattacharyya	**79,0323**	0,628	0,5884	79,0323	0,538	0,5373
ttest	77,4194	0,605	0,5607	**82,2581**	0,617	0,6164
wilcoxon	62,9032	0,311	0,2849	74,1935	0,436	0,4364

May be reached the following outcomes by referencing the above values;

✓ In both of datasets,the highest results of best 300 features were obtained by using Linear SVM classifier.

The ROC graphs of the above classification results are given below;

Figure V: *Ovarian dataset ROC graph (feature count "k" = 300)*

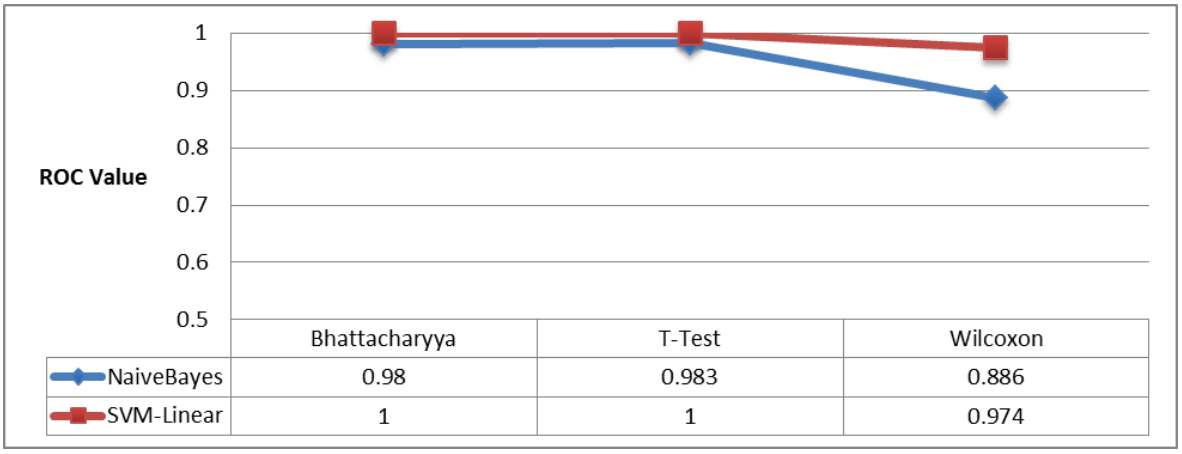
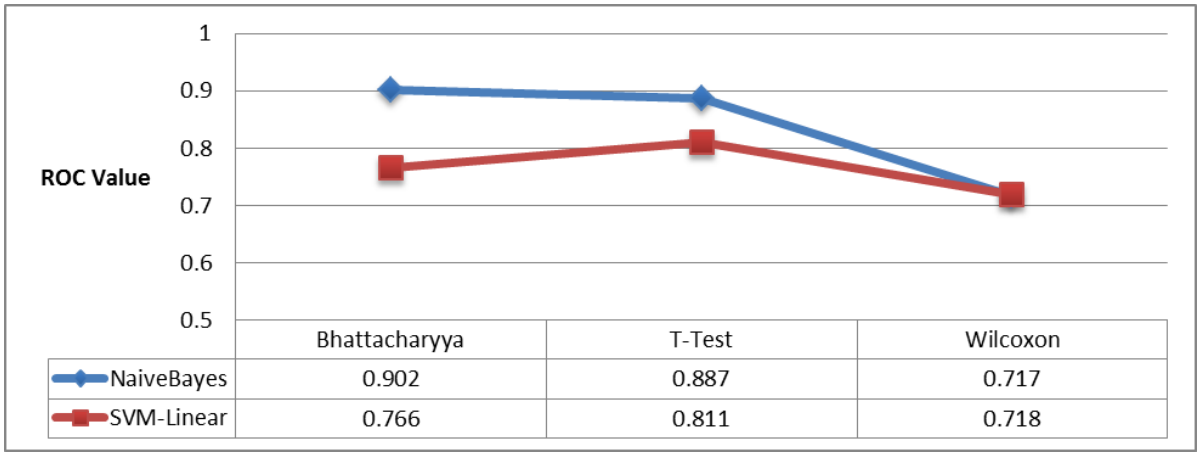


Figure VI: *Colon dataset ROC graph (feature count "k" = 300)*



III. CONCLUSION

"Average Accuracy Results Table" is formed with the average of the results which given in the above tables.The averagedtableis given below;

Table XII: *Average Accuracy Results Table ("k" is the number of features)*

Average Accuracy Results				
Datasets	***k***	***Bhattacharyya***	***T-Test***	***Wilcoxon***
Ovarian	***150***	**99,2095**	98,8145	93,8735
	300	98,2214	**98,4190**	90,3162
Colon	***150***	**80,6452**	78,2259	70,9677
	300	79,0323	**79,8388**	68,5484

Where,

The **green**cells show the highest average accuracy resultsof the Ovarian dataset.

The **blue**cells show the highest average accuracy results of theColon dataset.

Above table was created with the averaged results of all classifiers for each method.

Table XIII: *Average Accuracy Results Table ("k" is the number of features)*

k = 150 Accuracy Results (%)		
	Naive Bayes	**Linear SVM**
Wilcoxon (Ovarian)	88,5375	99,2095

$$Wilcoxon\ (avg) = \overline{88{,}5375 + 99{,}2095} = 93{,}8735 \quad (12)$$

Following conclusions are reached when considering the obtained average accuracy results

- Ranked Colon dataset results has been increased in comparison with raw dataset results. Hence, ranking-selection algorithms are quite useful for this dataset.
- Ranked Ovarian dataset results has been decreased a little in comparison with raw dataset results.Hence, ranking-selection algorithms is useful for the purpose of shorten the classification duration.
- Also, the effect of the Wilcoxon method was observed. This method is quite ineffective for all used datasets. Hence, it is not useful for these datasets.

At the next works; performance improvement can be realized with using same feature ranking algorithms and datasets. Also, new feature ranking methods can be used in the work.All processes can be repeated with less number of features. Roc and Accuracy values can be increased with using ensemble classifiers. Thus, the advantages and disadvantages of used each methods can be determined clearly.

IV. REFERENCES

Akbaş, A. *et al.* (2013) Performance Improvement with Combining Multiple Approaches to Diagnosis of Thyroid Cancer. *The 7th International Conference on Bioinformatics and Biomedical Engineering (iCBBE 2013)*, Beijing, China.

Alon, U. *et al.* (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl Acad. Sci. USA*,**96,** 6745–6750.

American Cancer Society (2014). *Cancer Facts & Figures*. Retrieved from http://www.cancer.org/research/cancerfactsstatistics/cancerfactsfigures2014/index

Babur, S. *et al.* (2012) Dvm Tabanlı Kalın Bağırsak Kanseri Tanısı İçin Performans Geliştirme. *Eleco 2012 Elektrik-Elektronik ve Bilgisayar Mühendisliği Sempozyumu*, 425-428.

Boser, B. E. *et al.* (1992) A training algorithm for optimal margin classifiers. *Proceedings of the fifth annual workshop on Computational learning theory(COLT)*, 144.

Cancer Research UK (2011, May 11) How many different types of cancer are there? *CancerHelp UK.*

Chee, M. *et al.* (1996) Assessing genetic information with high-density dna arrays. *Science*,**274**, 610–614.

Cohen, J. (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement,* **20** (1), 37-46.

Djouadi, A. *et al.* (1990) The quality of Training-Sample estimates of the Bhattacharyya coefficient. *IEEE Transactions on Pattern Analysis and Machine Intelligence.***12** (1), 92–97.

Golub, T.R. *et al.* (1999) Molecular classification of cancer: class discovery and class predition by gene expression monitoring. *Science*. **286,** 531–537.

Jemal, A. *et al.* (2011) Global cancer statistics. *CA: a cancer journal for clinicians,***61** (2), 69–90. doi:10.3322/caac.20107. PMID 21296855.

Jiaxi, L. (2010) The Application and Research of T-test in Medicine. *Networking and Distributed Computing (ICNDC).*

Loris, N. *et al.* (2012) Combining multiple approaches for gene microarray classification. *Oxford University Press,* **28** (8), 1151-1157.

Mann, H.B. and Whitney, D.R. (1947) On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, **18**, 50-60.

Matthews, B. W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)* - Protein Structure **405** (2), 442–451.

McCue, R. (2009) A Comparison of the Accuracy of Support Vector Machine and Naive Bayes Algorithms In Spam Classification. *University of California at Santa Cruz*, Nov 29.

Petricoin, E.F. *et al.* (2002) Use of proteomic patterns in serum to identify ovarian cancer. *Lancet*, **359,** 572–577.

Swets, A. (1996) Signal detection theory and ROC analysis in psychology and diagnostics. *Lawrence Erlbaum Associates*, Mahwah, NJ.

Taylor, R. (1999) An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements. 128–129.

Turhal, U. *et al.* (2013) Performance Improvement for Diagnosis of Colon Cancer by Using Ensemble Classification Methods. *The International Conference on Technological Advances in Electrical, Electronics and Computer Engineering (TAEECE 2013),* Konya, Turkey.

Wilcoxon, F. (1945) Individual comparisons by ranking methods. *Biometrics Bulletin,***1,** 80-83.

William, H. *et al.* (2007) *Support Vector Machines Numerical Recipes: The Art of Scientific Computing (3rd ed.).* Cambridge University Press, New York.

Zhang, H. (2005) Exlporing Conditions for the Optimality of Naive Bayes. *International Journal of Pattern Recognition and Artificial Intelligence,***19** (2), 183-192.

Uğur TURHAL was born in Trabzon, Turkey in 1988. He was graduated with Bachelor's degree from Marmara University in 2011. He is a graduate student in the Computer Engineering Department of Yalova University, Turkey. Also, He is working as a computer specialist at Balikesir University, Turkey. Interested areas are; Bioinformatics, Signal Processing, Microarray Datasets, Cancer Dieseases

Murat GÖK performed a Master in Computer Sciences at Mugla University (Turkey). After his Master thesis on the decision support systems, he began in 2006 a PhD in Computer Sciences at Sakarya University (Turkey). In June 2011, he defended his PhD thesis entitled "Prediction of HIV-1 Protese Cleavage Sites with New Techniques". Having completed his PhD, he became an assistant professor at the department of computer engineering on Yalova University(Turkey).His research interests are bioinformatics, machine learning algorithms and theories, computer programming. He has several papers on bioinformatics. He currently has several master students.

Suat ONUR was born in Kütahya, Turkey in 1972. He was graduated with Bachelor's degree from Gazi University in 1995. He is a graduate student in the Electric-Electronic Engineering Department of Balikesir University, Turkey. Also, He is working as a lecturer at Balikesir University, Turkey. Interested areas are; Bioinformatics, Internet Programming, Embedded Systems

Sebahattin BABUR was born in Bursa/Turkey in 1988. He was graduated with Bachelor degree in 2011 from Marmara University. He is a graduate student in the Computer Engineering Department of Yalova University, Turkey. Also, he has been working as Technical Sales Engineer for 1 year at Beckhoff Automation Company, Turkey. His areas of interest: Solution of Bioinformatics Problems, Image Processing, The Design of Electronic Circuits, Industrial Automation Technology