IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 7, 2019, accepted September 22, 2019, date of publication September 27, 2019,
date of current version October 9, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2944295

On the Feature Selection Methods and Reject Option Classifiers for Robust Cancer Prediction

MUHAMMAD HAMMAD WASEEM[1], MALIK SAJJAD AHMED NADEEM[1], ASSAD ABBAS[2], ALIYA SHAHEEN[3], WAJID AZIZ[1,4], ADEEL ANJUM[2], UMAR MANZOOR[5], MUHAMMAD A. BALUBAID[6], AND SEONG-O SHIM[4]

[1]Department of Computer Science and Information Technology, University of Azad Jammu and Kashmir, Muzaffarabad 13100, Pakistan
[2]Department of Computer Science, COMSATS University Islamabad, Islamabad 45550, Pakistan
[3]Department of Mathematics, University of Azad Jammu and Kashmir, Muzaffarabad 13100, Pakistan
[4]College of Computer Sciences and Engineering, University of Jeddah, Jeddah 23890, Saudi Arabia
[5]Computer Science Department, Tulane University, New Orleans, LA 70118, USA
[6]Industrial Engineering Department, King Abdulaziz University, Jeddah 21589, Saudi Arabia

Corresponding author: Assad Abbas (assadabbas@comsats.edu.pk)

ABSTRACT Cancer is the second leading cause of mortality across the globe. Approximately 9.6 million people are estimated to have died due to cancer disease in 2019. Accurate and early prediction of cancer can assist healthcare professionals to devise timely therapeutic innervations to control sufferings and the risk of mortality. Generally, a machine learning (ML) based predictive system in healthcare uses data (genetic profile or clinical parameters) and learning algorithms to predict target values for cancer detection. However, optimization of predictive accuracy is an important endeavor for accurate decision making. Reject Option (RO) classifiers have been used to improve the predictive accuracy of classifiers for cancer like complex problems. In a gene profile all of the features are not important and should be shaved off. ML offers different techniques with their own methodology for feature selection (FS) and the classification results are dependent on the datasets each having its own distribution and features. Therefore, both FS methods and ML algorithms with RO need to be considered for robust classification. The main objective of this study is to optimize three parameters (learning algorithm, FS method and rejection rate) for robust cancer prediction rather than considering two traditional parameters (learning algorithm and rejection rate). The analysis of different FS methods (including t-test, Las Vegas Filter (LVF), Relief, and Information Gain (IG)) and RO classifiers on different rejection thresholds is performed to investigate the robust predictability of cancer. The three cancer datasets (Colon cancer, Leukemia and Breast cancer) were reduced using different FS methods and each of them were used to analyze the predictability of cancer using different RO classifiers. The results reveal that for each dataset predictive accuracies of RO classifiers were different for different FS methods. The findings based on proposed scheme indicate that, the ML algorithms along with their dependence on suitable FS methods need to be taken into consideration for accurate prediction.

INDEX TERMS Cancer, classification, feature selection, genetic profile, machine learning, reject option.

I. INTRODUCTION

Cancer is becoming one of the main causes of death across the globe [1]. Approximately 9.6 million people are estimated to have died due to different types of cancer in 2019. Thousands of people die and agonize across the world every year due to inaccuracies in the healthcare systems. Genetic profile contains valuable information about the genes regulating cell growth and abnormalities occurring on the development of some specific cancer [2]. The extraction of this valuable information can assist in the reliable prediction of disease onset and in devising managerial solutions for the selection of therapy and personalized care [2]. The quality of prediction for effective decision making is of primary importance and therefore the main emphasis of science is assisting the insufficiencies of human findings and judgments [3]. In the fields of artificial intelligence, information science etc. numerous tools and techniques called decisions support systems (DSS)

The associate editor coordinating the review of this manuscript and approving it for publication was Vlad Diaconita.

have been developed for complex decision making. The idea of DSS is tremendously wide and its characterization varies with author's opinion [3]. In numerous areas namely medicine, business, military etc., DSSs are attaining good reputation. They are particularly appreciated in the circumstances where the accuracy and optimality are significant as well as where the total presented material is excessive for the perception of an unassisted human decision. While giving quick access to related information as well as assisting the procedure of constructing conclusions, DSS can assist human rational absences by assimilating several bases of knowledge. Such techniques may be adapted for rational decision in medical domain.

With the advances in technology, computers have progressed a lot in storing and retrieving large amount of data and have efficient access of data from remote location with great accuracy. In some scenarios, data is labeled which helps in categorization of a specific instance. In medical domain Gene Expression (GE) microarray data is considered one of the main sources of information which can be more accurately used to extract useful information and to build robust predictive systems. The main issue with GE microarray data is the curse of dimensionality (having thousands of features but few samples) that may be overcome by different feature selection (FS) methods including t-test [4], Las Vegas Filter (LVF) [5], Information Gain (IG) [6], and relief [7]. After reducing the number of features and having only relevant data different supervised ML algorithms linear discriminant analysis (LDA) [8], support vector machine (SVM) [9], random forest (RF) [10], k-nearest neighbors (kNN) [11] have been applied with certain degree of success but the use of both of these techniques still require improvement for robust decision making. To overcome the problem of low accuracy, RO (refraining from making decisions in case of ambiguity) is one of the methods proposed in ML literature [12]. ML algorithms use different mechanisms to improve their accuracy by incorporating the RO and some get more improvement than others [13]. Accuracy-Rejection Curves (ARCs) are used to compare the accuracy of RO classifiers in different rejection regions [13]. Previously, researcher used ARCs to compare the accuracies of different supervised ML algorithm without exploring the aspect of using different FS methods for more accurate models. We developed a method for simultaneously comparing the performances of classifiers in terms of their rejection rates (namely the RO classifiers), based on accuracy-rejection curves (ARCs) while selecting more relevant features by different FS methods for robust cancer prediction. We assume that, for a given sample, although rejection has different impacts on the accuracy of different classifiers, the performance of classifier also depends upon the FS method used. Therefore, robust RO classifier depends upon the FS method used. The objective of this study is to use various FS methods along-with different ML algorithms and varying RO settings for achieving robust accuracy. The purposed methodology of DSS using RO has been applied on publicly available datasets of colon cancer, leukemia and breast cancer. Analysis of empirical results shows that the selection of robust RO classifiers and FS method depends on dataset to be used and acceptable rejection rate.

In this section, significance of RO based ML classifiers and FS methods for accurate DSSs in medical field are presented. In section II, literature review and background of feature selection, RO classification, and significance of accuracy with different rejection regions are discussed. Section III demonstrates the operational details of the presented work, including definition and explanation of different RO classifiers and their working along with FS methods and the brief summary of datasets used in this study. Results obtained on the basis of experimental design (Section III) are presented and discussed in section IV. At the end general discussion and future plans are explained.

II. LITERATURE REVIEW

The magnitude of the data used in ML and data mining has drastically improved. Due to the huge number of features, a learning model leads to over fit and may result in decline of performance. To discourse the problem of high dimensionality, different techniques of dimensionality reduction have been studied in fields of ML and data mining including FS. In case of feature selection, a subset of features is chosen from the original feature space without any alteration and also maintaining the significance of original features. FS is extensively working aspect to decrease dimensionality of feature space by removing irrelevant features and consider reduced data for the use in classification tasks.

FS methods are classified into supervised [14], [15], unsupervised [16], [17] and semi-supervised categories [18], [19] based on the labeled and unlabeled training sets. Filter, Wrapper and embedded models are further classification of supervised FS techniques [20]. In this study, we are mainly concerned with filter methods i.e. t-test, LVF and IG.

The t-test [4] is most commonly used method for ranking genes by using t-value. In a study, the researchers used t-test to measure the class probability of genes for binary class problems [21]. Another study presented a comparison of five FS methods using two cancerous datasets [22].

Information Gain [6] has been used by for gene selection [22], [23], to measure the information by knowing the dependence between class label and feature value.

Another filter method based on probabilistic algorithm was proposed by [5] is typically known as Las Vegas Filter (LVF), which selects feature subset randomly from the feature space and fulfill the task.

After the predictive features are assembled using feature selection, the next step is to classify samples into classes. For this purpose, the classifiers that have been used in this study include linear discriminant analysis (LDA) [8], support vector machine (SVM) [9], random forest (RF) [10] and k-nearest neighbors (kNN) [11].

SVM are progressively becoming popular classifiers in diverse fields and has been widely used to classify GE data [21]–[23]. Naturally, SVM scans for a hyperplane

for separating data of two classes within margins. References [24]–[26] used SVM to the cancer classification problems and found that SVM has the highest accuracy on the cancerous data sets. Li *et al.* [27] also used SVM in their studies for tissue classification based on GE data.

kNN [28] method is based on distance functions like Euclidean distance, which classify samples according to the class of its k-nearest neighbor. In many applications, kNN has shown better performance than other complex methods [29], [30]. The kNN approach has been applied for gene classification by measuring the similarity between pairs of samples [31].

LDA is another method introduced by R. A. Fisher in 1936 for the classification purpose [8]. This technique is used in different fields (ML, pattern recognition) and provides a model with good accuracy for finding linear arrangements of samples to separate them into two or more classes. LDA is widely used in GE microarray data analysis [31] and for the cancer classification problem [31].

Random forest (RF) classification algorithm was developed by Breiman [10]. This is a classification method based on ensemble learning, which builds a chain of classification trees by using random samples from original samples. RF owns various properties which makes it attractive for microarray gene expression data classification [32]. Diaz-Uriarte and de Andres [33] evaluated the use of RF in classification of GE microarray data and concluded that RF classifiers performed better classification as compared to other classifiers in GE microarray data. RF was also explored for prediction and classification in medical domain by [34]–[36]. SVM, RF, LDA and kNN algorithms have been used by [30], [37]–[41] for classification and prediction purposes in different domains.

Reject option (RO) technique can be used when the classifier is not adequately precise for the job at hand. The RO presented by [12], suggested that in order to decrease the probability of error, the samples that have inefficiently high posteriori likelihoods shouldn't be classified. In the feature space the rejection area is well-defined and all the instances that lie in this area are rejected. If the prediction is not satisfactorily consistent and it gets into the rejection area, then the classifier rejects an instance. According to [42], there exists an inverse correlation between rejection rate and error rate as with the increase of rejection rate the error rate decreases. The basic parameters in the classifiers with RO are thresholds and these thresholds describe the rejection areas. References [43]–[45] and other researchers have suggested various approaches for describing an optimum rejection rule. In this work, we computed accuracy against different reject thresholds.

The present study uses the method of ARCs [13] in order to compare the performance of various RO classifiers after reducing data using different FS methods in terms of diverse rejection rates. All ARCs start form a point $(0, a)$, where 0 means 0% rejection and a is the accuracy percentage of the classifier. ARCs converge on the point (1, 1) because the accuracies of the classifiers are 100% for 100% rejection rate. ARCs are quite beneficial for graphical comparison of the accuracy of classifiers as a function of their rejection rates.

In various fields of medical research microarrays are used, as they provide instantaneous expression assessments for thousands of genes. Prediction of the biological factors that are based on the gene-expression profile is the ultimate capable application. These expression profiles can be used to distinguish various types of tumors and ultimately can be helpful for selecting the suitable therapeutic intervention.

III. MATERIALS AND METHODS

A. DATASETS

In the present study three GE microarray datasets were used. Leukemia dataset [46] comprises of 72 patients from which 47 patients with acute lymphoblastic leukemia (ALL) and 25 patients with acute myeloid leukemia (AML). The Colon cancer dataset [47] contains the genetic profile of 39 patients who are affected by colon cancer and 23 non-affected (healthy) subjects. The Breast cancer dataset [48] contains data of 295 patients that were suffering from breast cancer, from which 115 have a good-prediction class whereas 180 have the poor-prediction class.

B. METHODOLOGY

Machine learning process in cross validation and FS settings is characterized by a series of steps as illustrated in the Fig. 1. Generally, data is defined by the relation

$$\chi = \{x^r, y^r\}_{r=1}^{N} \tag{1}$$

where x represents the feature space, y represents the target, r represents example in the dataset χ.

Target may be categorical values (in case of classification problem) or may be continuous values (in case of regression problem).

To have a generalized predictive model, data is divided into k-folds (Fig. 1 section Cross Validation). After creating folds, train and test sets are obtained using the notion of CV [18]. Then, the list of relevant features (computed from train data) using FS methods (t-test, LVF, IG and relief) is obtained. This list is then used to reduce the Train data and test data (as depicted in Fig. 1 section Feature Selection). Training data is used to build the predictive model while test set is used to evaluate the performance of built model as shown in the Fig. 1 (section Training and Testing). ARCs (Fig.1 Section Final Results) are generated using the methodology of [13].

Generally, a dataset may have irrelevant features, therefore it is wise to reduce the data into smaller subsets based on some feature selection method [14], [15]. In the perspective of classification, FS methods can be categorized in to three groups (filter, wrapper and embedded) depending upon the fact that how these attribute identification methods will combine with the creation of classification model. Feature selection techniques shorten the training time, improves generalization and model interpretability without losing the importance of single attribute.

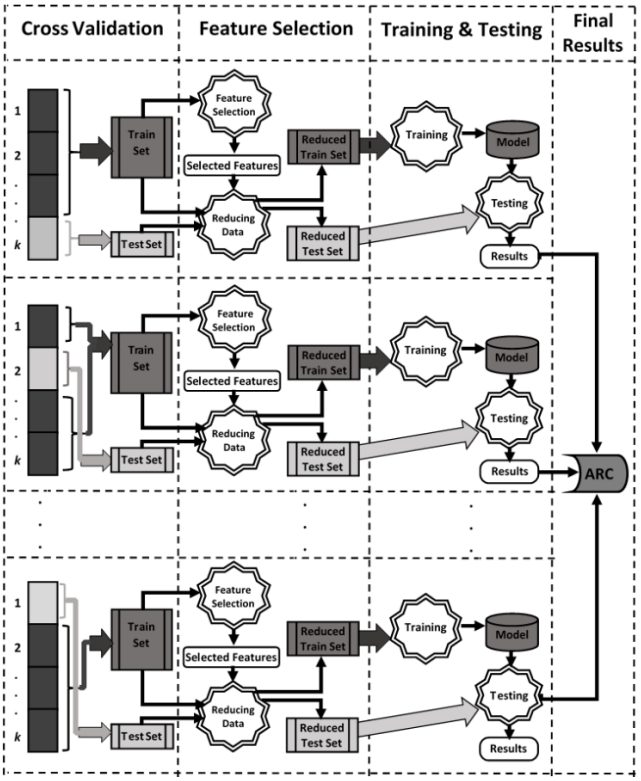


FIGURE 1. k-fold CV based ML process.

Given a feature set χ (as defined in equation 1) having m features, a feature selection method (FS) selects a subset χ_s having s more relevant features ($s < m$) based on the criteria laid down in the feature selection method,

$$\chi_s = arg \max_s FSMethod\{x_i | i = 1 \ldots m\} \tag{2}$$

After reducing the dimensions of training and test sets using equation 2, different models are learned using learning algorithms discussed earlier, each producing a classifier. The purpose of the classifier is to assign a discrete class label (from one of the target values among A, B, C, ... N) to unseen samples i.e.

$$f(\chi) = y = \begin{cases} A \\ B \\ C \\ \vdots \\ N \end{cases} \tag{3}$$

whereas in case of binary classification the target is limited to 2 discrete class labels and hence equation 3 implies

$$f(\chi) = y = \begin{cases} 1 & if\ x\ \in +ive\ class \\ 0 & if\ x\ \in -ive\ class \end{cases} \tag{4}$$

Mostly a classifier's capability is assessed through its error rate. When the classifier is not adequately precise for the job at hand, reject option (RO) technique [12] can be used. The basic parameters in the classifiers with RO are thresholds and these thresholds describes the rejection areas. Therefore, in case of RO classifiers, rejection rate is also considered along with error rate to assess the classifier's performance.

According to [42] a sample x is accepted only if the probability that x belongs to class y_i is higher than or equal to the threshold t otherwise the prediction is not reliable and the classifier should reject the sample. Classifier accuracy (CA) for a binary classification problem (equation 4) is defined by following relation

$$CA(x) = \begin{cases} argmax\ y_i(p(y_i \mid x) & if\ |(p(C_1 \mid x) - p(C_2 \mid x))| \geq t \\ reject & if\ |p(C_1 \mid x) - p(C_2 \mid x)| < t \end{cases} \tag{5}$$

According to the best of our knowledge, no method or model has yet been proposed to select a robust FS method out of available/used FS methods and a RO classifier obtained using a range of rejection thresholds. In this work, we have proposed to use following relation (using equations 2, 4, & 5), for simultaneous selection of robust RO classifier (*RROCls*) and FS method for the classification or prediction problem under consideration

$$RROCls = arg \max_{i,j,rT} g(CA_i(\chi_s) | (\chi_{s_j}, t_{rT})) \tag{6}$$

where CA_i represents the accuracies of RO classifiers, χ_{s_j} are the features selected by one of j FS methods and $0.0 < t_{rT} < 1.0$ is the range of rejection thresholds.

The relation presented in equation 6 can be helpful for exploring robust RO classifier and FS method using different values of rejection thresholds t_{rT}.

The complete process is summarized in the algorithm I. According to the best of our knowledge, before this study, no algorithm has yet been proposed to simultaneously select the optimal FS method and RO classifier selection for a user preferred criterion (accuracy). In this section, the proposed algorithm to select optimal FS method and RO classifier for robust predictability of cancer is presented, which is the main contribution of the current study. The algorithm starts by taking a matrix of original data ($\chi = \{x^r, y^r\}_{r=1}^N$) where x^r is the feature space, y^r the target, and r represents an example in the dataset χ. *FSms* is a user provided list of FS methods, *TLAlgo*: a list of traditional learning algorithms without RO. *nbrBF*: is the number of best features to be selected, *CriL* parameter shows minimum desired accuracy for robust FS method and RO classifier selection. The algorithm returns $RROCLs_{FSms}$ which shows the optimal FS method and RO classifier among the provided *FSms* and *TLAlgos* as arguments to the algorithms for robust cancer prediction according to the details laid down in the algorithm and Eq. 06.

The algorithm splits the original data $\chi_{k=1}^n$ into k distinct folds (D_k) as per line 1. In lines 3 and 4, by considering all the k folds, alternatively each of the folds is used as test set (ts_k) while rest of the folds are taken as train set (tr_k). For each of the FS methods ($FSms_l$), the algorithm from line 6 to 11, first finds the *nbrBF* best features from train set (line 6), then

Algorithm 1

Input: $(\chi, FSms, nbrFS, TLAlgo, CriL)$ where

- $\chi = \{x^r, y^r\}_{r=1}^{N}$ is a matrix of original data where x^r represents the feature space, y^r the target, while r denotes an example in the dataset χ.
- $FSms$; a list of feature selection methods
- $TLAlgo$; a list of traditional learning algorithms without RO.
- $nbrBF$; number of best features to be selected
- $CriL$; a list of parameters (desired Accuracy, acceptable Rejection Rate) for robust RO classifier selection

Output: Robust RO Classifier ($RROCls_{FSms}$) using Eq. 06 based on criterion set by the user as a $CriL$ argument

1. Split data χ into k folds i.e. $D_k = \chi_{k=1}^{n}$
2. **Repeat** steps 3 to 12 for each fold k, where $k := 1$ to n
3. $tr_k := D_{n-1}$ ▷ train set
4. $ts_k := D_n$ ▷ test set
5. **Repeat** steps 6 to 12 for each feature selection methods ($\boldsymbol{FSms_l}$) where $l := 1, 2, 3, \ldots, j$
6. $BestFeatures_{k,FSms_l} := FSMethod_l(tr_k, nbrBF)$
7. $Reduced\_tr_{k,FSms_l} := tr_k[BestFeatures_{k,FSms_l}]$
8. $Reduced\_ts_{k,FSms_l} := ts_k[BestFeatures_{k,FSms_l}]$
9. **Repeat** steps 10 to 11 for each $TLAlgo_c$ where $c := 1, 2, 3, \ldots, t$
10. $Model_{k,FSms_l,TAlgo_c} := BuildClassifier(Reduced\_tr_{k,FSms_l}, TAlgo_c)$
11. $ROCls\_Result_{k,FSms_l,TAlgo_c} := TestModel\left(Model_{k,FSms_l,TAlgo_c}, Reduced\_ts_{k,FSms_l}\right)$
12. **Repeat** steps 13 to 15 for each $TLAlgo_c$ where $c := (1, 2, 3, \ldots, t)$
13. **Repeat** steps 14 to 15 for each $FSms_l$ where $l := 1, 2, 3, \ldots, j$
14. **Repeat** step 15 for each $rejT$ where $rejT :=$ $1, 2, 3, \ldots, 100$ ▷ $rejT$ is the rejection threshold
15. $Avg\_Accu\_ARC\_ROCls[c, FSms_l, rejT] :=$ $avgARC(ROCls\_Result_{\bigcup_1^n k,FSms_l,TAlgo_c}, rejT)$
16. $RROCls_{FSms} := Avg\_ARC\_ROCls[1, 1, CriL]$
17. **Repeat** steps 19 to 21 for each $TLAlgo_c$ where $c := (2, 3, \ldots, t)$
18. **Repeat** steps 20 to 21 for each $FSms_l$ where $l := 2, 3, \ldots, j$
19. if $(Avg\_Accu\_ARC\_ROCls[c, FSms_l, CriL] > RROCls_{FSms})$
20. $RROCls_{FSms} := Avg\_Accu\_ARC\_ROCls[c, FSms_l, CriL]$
21. **Return** $RROCls_{FSms}$

reduces the train and test sets (lines 7 & 8) by considering only the $nbrBF$ as chosen in line 6. Then, for each of $FSms_l$ and $TLAlgo_c$, predictive model is built (line 10) and newly built model is tested (line 11). For each combination of traditional learning algorithms from $TLAlgo$ and FS methods from $FSms_l$, mean values of accuracies $Avg\_Accu\_ARC\_ROCls$ against every rejection threshold $rejT$ from all k folds is computed to have average $ARCs$ of $ROCls$ used in the algorithm (please refer lines 12 to 15). Finally, from these average $ARCs$ and as per the criterion of acceptable rejection rate ($CriL$), a robust RO classifier and a FS method ($RROCls_{FSms}$) are obtained and returned using lines 16 to 21.

IV. RESULTS

In this section the results obtained are presented and compared with the previously published work by Nadeem *et al.* [13] which is the only study that uses accuracy rejection curves (ARCs) to compare the performances of different learning algorithms. In their study they only used t-test FS method on above mentioned datasets. To validate and verify the hypothesis of the present study besides t-test FS method, we have used LVF, IG and relief to get robust results with respect to accuracy while checking the variations on different RO classifiers. Moreover, we also compared the obtained results with those of Yeh *et al.* [49]. They used two FSMs (IG and t-test) and only DTs as learning algorithm without considering RO.

Results obtained by performing classification using proposed simulations settings (as illustrated in Fig. 1 and Equations 2 and 5) are presented in term of ARCs. Classification has been done at different rejection rates ranging from 0% to 100% (where 0% shows no rejection and 100% being the maximum rejection rate) for three different cancerous (colon, leukemia and breast cancer) datasets. Five classifiers which include LDA, kNN, SVM (Linear and Radial kernels) and RF are used to build predictive model along with four different feature selection methods (LVF, IG, t-test and Relief). Results are compared FS method wise where each FS method is compared using all classifiers for all datasets separately.

In Fig. 2 ([A], [B], [C], [D]), results obtained using different FS methods against different RO classifiers separately with colon dataset are shown. It is clear from the figure that

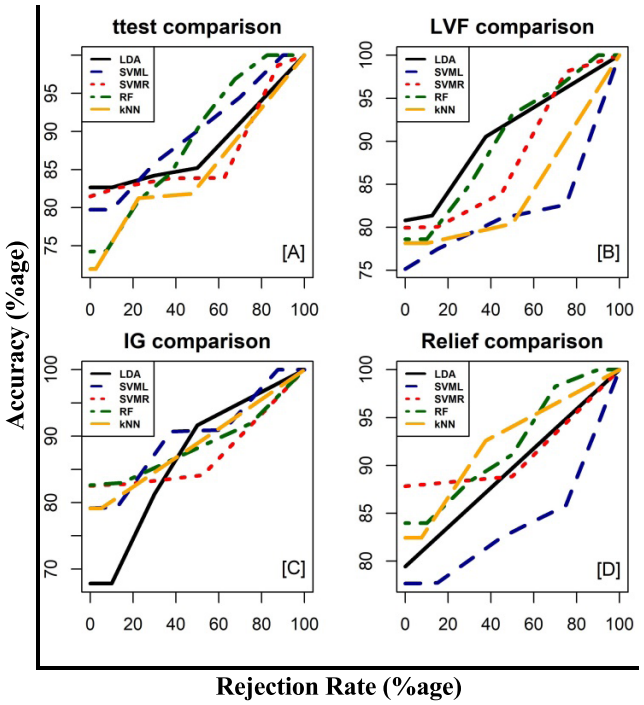


FIGURE 2. [A] ttest, [B] LVF, [C] IG, [D] Relief and ARC's of different feature selection method in combination with different classifiers using Colon Dataset [47].

without rejection (WoR), Relief with SVM-R has 87.5% accuracy, t-test with LDA (83%), and LVF with LDA (81%). With rejection, classification accuracy of RF using three out of four FS methods used in this study is higher as compared to all other classifiers at the same rejection rate. Whereas in case of IG, without rejection, RF classifier gives better separation (83%), and with rejection SVM-L perform better among all classifiers (82% rejection rate).

Contrary to this approach, if we compare the performances of RO classifiers solely using one FS method (t-test for example) as was done by Nadeem *et al.* [13] we have RF 86% accuracy at 40% RR. Whereas on the same RR, LVF with LDA has 91.5% accuracy, IG with kNN is 91% accurate while relief with kNN is 93.5% accurate. The overall results depict that it is better to perform a comparative analysis of available FS methods and RO classifiers for robust classification of cancer as was hypothesized before.

The results presented in Fig.2 are summarized in table 1. Here accuracy is computed at discrete rejection rates up-to 50%RR against FS methods (FSM) and learning algorithms (learner) used to build the predictive models.

If we look at ARCs presented in Fig. 2 and results shown in table 1, classifiers performance continuously grows to the maximum accuracy with the increase in rejection rate.

Table 1 shows that WoR accuracy of the LDA classifier using t-test and LVF FS methods is highest compared to the other classifiers. Whereas with the increase in rejection, the accuracy of classifiers built using other learning algorithms and FS methods also increase. If we look at 40% RR in case of t-test method, SVML classifier gives highest accuracy than others and on 50 % RR in case of t-test and LVF, RF classifier perform better as compared to all other classifiers and have highest accuracy among others. Similar results are obtained in the case of other FS methods, that WoR, classifiers are not performing better ends up with higher accuracy with the increase in rejection rate (table 1). Overall results depict that better classification for the colon dataset can be achieved with reject option.

TABLE 1. Accuracy of different classifiers on different rejection rate using fs methods for colon cancer dataset.

FSM	Learner	ACCURACY						
		WoR	5% RR	10% RR	20% RR	30% RR	40% RR	50% RR
Ttest	LDA	83.0	83.0	83.0	84.0	84.0	84.5	85.0
	SVML	79.0	81.0	79.5	85.0	86.0	87.0	91.0
	SVMR	82.0	83.0	81.5	83.0	83.0	83.5	84.0
	RF	74.0	76.5	74.0	82.5	84.0	85.5	92.0
	KNN	72.0	74.0	72.0	81.0	81.5	82.0	85.0
IG	LDA	68.0	68.0	68.0	74.5	82.5	91.0	91.5
	SVML	79.0	79.0	79.5	82.0	88.0	90.0	92.0
	SVMR	83.0	83.0	82.5	83.0	83.0	83.5	84.0
	RF	83.0	83.5	84.0	84.5	85.0	86.0	86.5
	KNN	79.0	80.5	81.0	83.0	85.0	86.5	87.0
LVF	LDA	81.0	82.0	82.0	84.0	91.0	91.0	93.0
	SVML	75.0	76.0	76.5	77.0	79.0	80.0	81.5
	SVMR	80.0	79.5	79.0	80.5	82.0	83.5	85.5
	RF	78.5	78.0	78.0	82.0	85.0	88.0	94.0
	KNN	78.0	78.0	78.0	79.0	79.0	79.0	80.0
Relief	LDA	79.5	80.5	81.5	83.0	76.5	87.0	89.0
	SVML	77.5	78.0	78.0	79.0	80.5	82.0	83.5
	SVMR	87.5	87.0	87.0	87.5	87.0	87.0	88.5
	RF	84.0	85.0	86.0	87.0	89.0	90.0	91.5
	KNN	82.0	82.5	85.0	87.0	92.0	93.5	94.5

Fig. 3 ([A], [B], [C], [D]) illustrate the results obtained using different feature selection methods against different classifiers separately for Leukemia cancer dataset. It is obvious from the figure that without rejection in case of Relief with SVM-R and t-test with RF has classification accuracy 88% and 92.5% respectively, which is improved with rejection. Without rejection, RF classifier gives better accuracy (82.5%), and with rejection SVM-L provides better results for IG FS method. In case of LVF, LDA provides better accuracy with and without rejection.

In some cases, it may also be seen that the classifier which performs better than other classifiers at low RRs, the increase in its performance becomes less when compared to other classifiers with the increase in RR. It is also possible that the classifier which was not performing well at low RR gives better accuracy with the increase in RR when compared to its competing classifiers. As shown in Fig. 3 [A], the SVM-R lacks in accuracy without rejection but with increase in rejection rate its accuracy also increases and on 60% RR it outclasses all other classifiers and gives best classification accuracy.

Table 2 shows that WoR, accuracy of the LDA classifier using LVF, RF using t-test and IG and SVMR using Relief FS methods are the highest compared to all the other classifiers.

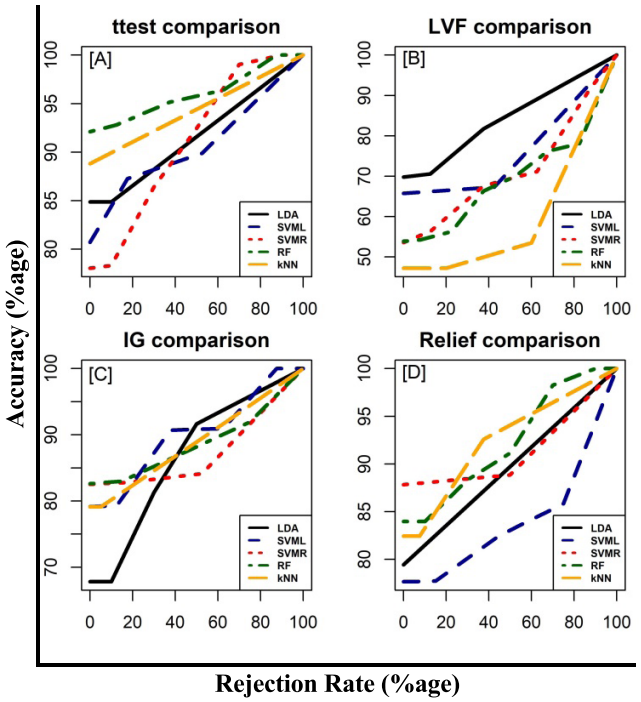


FIGURE 3. [A] ttest, [B] LVF, [C] IG, [D] Relief ARC's of different feature selection method in combination with different classifiers using Golub Dataset [46].

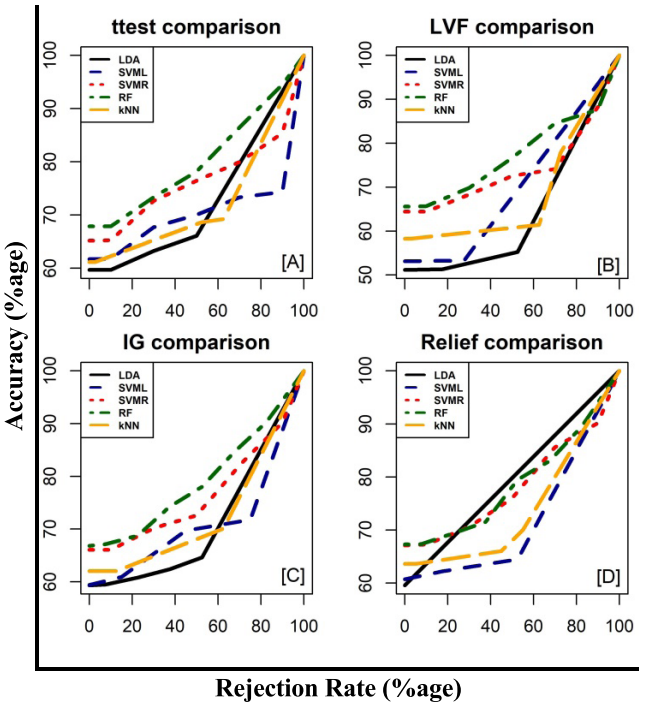


FIGURE 4. [A] ttest, [B] LVF, [C] IG, [D] Relief ARC's of different feature selection method in combination with different classifiers using Breast Cancer Dataset [48].

TABLE 2. Accuracy of different classifiers on different rejection rate using fs methods for Leukemia cancer dataset.

FSM	Learner	Accuracy						
		WoR	5% RR	10% RR	20% RR	30% RR	40% RR	50% RR
Ttest	LDA	85.0	84.5	85.5	86.5	88.0	89.0	91.0
	SVML	81.0	83.0	85.0	87.0	86.5	87.0	87.5
	SVMR	78.0	78.0	79.0	82.5	86.0	88.5	91.0
	RF	92.5	92.5	93.0	93.5	94.0	95.5	96.0
	KNN	89.0	89.5	90.0	91.0	91.5	92.0	94.5
IG	LDA	69.0	69.0	70.0	75.0	81.0	86.0	92.0
	SVML	79.5	79.5	80.0	83.5	89.0	90.5	92.5
	SVMR	83.0	83.0	83.0	83.5	82.0	82.5	83.0
	RF	83.0	83.5	84.0	84.5	86.0	86.5	87.0
	KNN	79.5	81.0	82.0	83.0	84.0	85.0	87.5
LVF	LDA	70.0	71.0	71.5	73.5	76.0	81.5	86.0
	SVML	67.0	67.0	67.5	67.0	67.0	67.0	69.5
	SVMR	52.0	55.0	58.0	60.0	64.0	68.0	69.0
	RF	52.0	52.0	53.0	54.0	62.0	68.0	69.0
	KNN	49.0	49.0	49.0	48.5	49.5	51.0	51.5
Relief	LDA	79.5	81.0	82.5	84.0	85.0	87.0	89.5
	SVML	78.0	78.5	78.0	79.0	81.0	82.0	83.5
	SVMR	88.0	87.0	88.0	88.5	88.5	88.0	88.5
	RF	84.0	85.0	86.0	87.0	89.0	90.0	91.5
	KNN	83.0	84.0	85.5	87.5	90.0	93.0	94.5

With the increase in RR, accuracy of the classifiers built using other learning algorithms and FS methods also increase. At 40% RR in case of t-test FS method, RF classifier gives highest accuracy than others and on 50 % RR in case of t-test, the RF classifier perform better as compared to all other classifiers and have highest accuracy among others. It may be the case that the classifier which perform better without rejection may performs better with rejection in some cases. E.g. at 30 % rejection, in Relief FSM, KNN performs better and remains higher accuracy at 40 and 50 % rejection rates. Same is in the case of other FSM, that WoR the classifiers which may not perform better ends up with higher accuracy with the increase in rejection rate as shown in table 2.

Results obtained by using different FS methods and classifiers for breast cancer dataset are shown in Fig. 4 ([A], [B], [C], [D]). It is clear from the figure that for all FS methods RF provides better classification accuracy with and without rejection. The accuracy is positively associated with rejection rate, i.e., increase in rejection rate also causes increase in accuracy

Table 3 shows that WoR the accuracy of the RF classifier using all FS methods is highest compared to all the other classifiers. Whereas with the increase in RR, the accuracy of classifiers built using other learning algorithms and FSM also increase. RF classifier performs better as compared to all other classifiers and has highest accuracy among others in all the FS methods expect Relief throughout the analysis.

For robust classification equation 6 may be used by setting value of k (acceptable rejection rate). Tuning k can give optimum results to help in making decisions, which is one of the main concerns in DSS and medical domain.

V. DISCUSSION

Predictive accuracy is the main concern for decision support systems especially in healthcare domains [50]. In healthcare, the development of GE microarray data-based decision support systems is the key area of research nowadays. Literature

TABLE 3. Accuracy of different classifiers on different rejection rate using fs methods for breast cancer dataset.

FSM	Learner	ACCURACY						
		WoR	5% RR	10% RR	20% RR	30% RR	40% RR	50% RR
ttest	LDA	60.0	60.5	61.0	61.5	62.0	63.0	65.0
	SVML	61.5	62.0	63.0	64.0	66.0	68.0	67.0
	SVMR	65.0	66.0	68.0	69.5	71.0	72.0	72.0
	RF	68.0	69.0	70.0	71.0	73.0	75.0	78.0
	KNN	61.0	62.0	63.0	63.5	64.0	64.5	66.0
IG	LDA	59.0	59.0	59.5	60.0	61.0	61.5	64.0
	SVML	59.0	60.5	61.0	62.0	67.0	68.5	69.0
	SVMR	66.0	65.5	68.0	69.5	71.0	72.0	78.0
	RF	67.0	68.0	69.5	70.0	73.0	77.0	79.0
	KNN	61.0	63.5	63.0	64.0	65.5	66.0	69.0
LVF	LDA	51.0	51.0	51.0	51.5	52.0	52.5	54.0
	SVML	52.0	52.0	52.0	52.0	53.0	61.0	66.5
	SVMR	65.0	66.0	68.0	69.0	69.5	70.0	70.0
	RF	66.0	67.0	69.0	70.0	71.0	71.5	75.0
	KNN	58.0	58.0	59.0	60.0	60.0	60.0	60.5
Relief	LDA	60.0	62.0	65.5	69.0	73.0	76.0	80.5
	SVML	60.5	61.0	62.0	63.0	63.0	64.0	64.0
	SVMR	68.0	69.0	69.5	70.0	73.0	73.0	78.0
	RF	68.0	69.0	69.5	70.0	70.5	71.0	79.0
	KNN	64.0	65.0	65.0	66.0	68.0	68.5	69.0

shows that the features of the samples are important in such systems. genetic profile of a person (healthy or diseased subject) may be of great importance for developing an efficient DSS. Moreover, genes are not only affected by the external environment, but expression values of certain genes can also be affected due to certain diseases like obesity, cancer etc. and the variations in these values are used to study the responses of different therapies.

For GE Microarray data based DSS, not all of the features are of importance and ML offers different techniques with their own methodology to select suitable features. Typically, in classification/ prediction, each given sample is assigned a class label without considering the degree of confidence on the classification which causes high error rates. ML literature shows that the use of reject option classifiers proved the predictive accuracy of classifiers for complex problems. Moreover, each dataset has its own distribution. In this work, the hypothesis is that FS methods and ML classifiers give diverse accuracies in changing reject option scenarios and with the change in datasets. Therefore it is wise to make an analysis of feature selection and ML algorithms for the selection of more suitable feature selection methods and ML algorithm for the problem under consideration.

In this study, we analyzed the use of various reject option classifiers with different feature selection methods along-with different ML algorithms. Different rejection rates are used for the comparison of each of the reject option classifier and feature selection methods. The analysis of GE Microarray dataset depicted that the accuracy of reject option classifiers improves by selecting features with different feature selection methods and by rejecting ambiguous predictions. In analysis of Colon cancer GE microarray dataset, RF classifier with relief FS method provides more accurate prediction than other FS methods and classifiers used in this study. In analysis of Leukemia cancer, the use of RF classifier with IG FS method shows better classification than the other RO classifiers.

TABLE 4. Comparison of accuracies obtained using 3 datasets (1: Colon Cancer, 2: Leukemia Cancer, 3: Breast Cancer) with previous studies.

Data used	Highest Accuracy achieve by		
	[49]	[13]	Current Study
1	ttest & DTs (77.42%)	ttest & LDA (91% @ 40%RR)	Relief & KNN (**93.5%** @40%RR)
2	ttest & DTs (87.22%)	ttest & SVMR (93% @ 40%RR)	ttest & RF (**95.5%**@40%RR)
3	Dataset not used	ttest & RF (73.5%@ 40%RR)	IG & RF (**77%**@40%RR)

In analysis of Breast cancer dataset, RF classifier is less accurate without rejection than the other classifiers. Here in case of RF with Reject Option improves its predictive capability by all four FS methods.

Table 4 summarizes the comparison of results obtained using three benchmark datasets in current study and the two of the previous studies by Nadeem *et al.* [13] and Yeh *et al.* [49]. Yeh *et al.* [49] used IG and ttest FSMs but only used DTs as learning algorithm. Nadeem *et al.* [13] used ttest as FSM, learning algorithms (LDA, SVMR, SVML, & RF), 10-fold CV and used reject option in their study. Although [13] and current study obtained accuracy at different rejection rates but just for comparison purposes only accuracies at 40%RR (which is less than the threshold of random guess i.e. 50%RR) are shown here. Current study uses four FSMs (ttest, LVF, IG, Relief), five learning algorithms (LDA, SVMR, SVML, RF & kNN), 3-fold CV and used reject option. Table 4 shows that current study achieves better accuracy by using different FSMs. Using Colon Cancer dataset [47], Yeh *et al.* [49] obtained 77.42% accuracy (using ttest FSM and DTs as learning algorithm), [13] got 91 % accuracy at 40% RR with ttest and LDA. Current study obtains 93.5% accuracy at 40% RR when Relief FSM is used along with kNN learning algorithm.

Row two of Table 4 shows the results with Leukemia cancer [46]. Here 87.22% accuracy is achieved by applying ttest and DTs as reported in [49] while [13] obtained 93% accuracy at 40% RR when they used ttest and SVMR. Current study obtains 95.5% accuracy at 40% RR when ttest FSM is used along with RF learning algorithm.

The third comparison Table 4 shows that [49] did not used Breast cancer dataset [48] in their study. Yet Nadeem *et al.* [13] achieved 73.5 % accuracy at 40% RR with ttest and RF. Current study obtains 77% accuracy at 40% RR when IG FSM is used along with RF learning algorithm.

This study and comparison in Table 4 reveals that using different FSMs along with different learning algorithms provides more options in terms of FSMs and learning algorithms for the selection of robust cancer classification. Although ttest FSM gives better results with Leukemia Cancer dataset [46] in all three studies listed in Table 4 but the results obtained from the other two datasets depict that the choice of FSM and

learning algorithm for robust cancer classification demands a comparison of different FSM and learning algorithms in RO scenarios.

VI. CONCLUSION AND FUTURE WORK DIRECTIONS

In this paper, we analyzed the dependence of RO based ML algorithms on FS methods for robust prediction of cancer while using GE microarray data. GE microarray data potentially have the so-called curse of dimensionality (having few samples but thousands of features) and generally, GE microarray data may have features which are not relevant to a problem under study. In such situations, according to Occam's razor [51], unnecessary information (redundant and irrelevant) should be shaved off to have a simpler predictive model. In a typical ML process, FS methods are used to remove the such unnecessary features. Each of traditional FS methods has its own mechanism to avoid irrelevant features. The idea of this study was to incorporate the FS methods to reduce the dimensionality of GE microarray data and RO classifiers in an algorithm to have robust cancer predictions in varying RO scenarios as discussed earlier. The results reveal that the RO classifiers showed improvements in predictive accuracy of RO classifiers differently at different Rejection rates and with the reduction of feature space with different FS methods. The presented analyses for cancer data show that predictive accuracy of different RO classifiers is subject to two parameters including opted FS method and the rejection rate. Therefore, it can be inferred that for a specific predictive problem it is optimal to make a comprehensive analysis of RO classifiers coupled with different FS methods at varying rejection rates. Moreover, ARCs are helpful in the selection of a robust RO classifier and FS method on the basis of customized requirements (maximum acceptable rejection rate and/or required accuracy). The selection of robust RO classifier and FS method may be of interest for physicians and surgeons in healthcare domain for optimal decision making. More accurate decisions may be helpful for effective and timely therapy of the patients,

This proposed methodology was used on publicly available GE microarray data. However, the same methodology may be explored for primary datasets and clinical parameters.

Although the results presented in this study show improvements in accuracy yet some aspects of FS methods and RO classifiers are to be explored. For example, there is always a cost associated with correct decisions, incorrect decisions and refraining from making a decision. Therefore, a cost-based study of different FS methods with RO classifiers may be of interest. Moreover, analyses of FS methods with bagging and boosting based RO classifiers may also be potential research directions.

REFERENCES

[1] J. Ferlay, I. Soerjomataram, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. M. Parkin, D. Forman, and F. Bray, "Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012," *Int. J. Cancer*, vol. 136, no. 5, pp. E359–E386, Mar. 2015.

[2] T. A. Manolio *et al.*, "Finding the missing heritability of complex diseases," *Nature*, vol. 461, pp. 747–753, Oct. 2009.

[3] M. J. Druzdzel and R. R. Flynn, "Decision support systems. Encyclopedia of library and information science. A. Kent," *Marcel Dekker, Inc. Last Login*, vol. 10, no. 3, p. 2010, 1999.

[4] B. L. Welch, "The generalization of 'student's' problem when several different population variances are involved," *Biometrika*, vol. 34, pp. 28–35, Jan. 1947.

[5] G. Brassard and P. Bratley, *Fundamentals of Algorithmics*. vol. 33. Upper Saddle River, NJ, USA: Prentice-Hall, 1996.

[6] T. M. Mitchell, *Machine Learning. WCB*. Boston, MA, USA: McGraw-Hill, 1997.

[7] K. Kira and L. A. Rendell, "The feature selection problem: Traditional methods and a new algorithm," presented at the Proc. 10th Nat. Conf. Artif. Intell., San Jose, CA, USA, 1992, pp. 129–134.

[8] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[9] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[10] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[11] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *Amer. Statist.*, vol. 46, no. 3, pp. 175–185, 1992.

[12] C.-K. Chow, "An optimum character recognition system using decision functions," *IRE Trans. Electron. Comput.*, vol. EC-6, no. 4, pp. 247–254, Dec. 1957.

[13] M. S. A. Nadeem, J.-D. Zucker, and B. Hanczar, "Accuracy-rejection curves (ARCs) for comparing classification methods with a reject option," in *Proc. Mach. Learn. Syst. Biol.*, 2010, pp. 65–81.

[14] L. Song, A. Smola, A. Gretton, K. M. Borgwardt, and J. Bedo, "Supervised feature selection via dependence estimation," in *Proc. 24th Int. Conf. Mach. Learn.*, Jun. 2007, pp. 823–830.

[15] J. Weston, A. Elisseeff, B. Schölkopf, and M. Tipping, "Use of the zero-norm with linear models and kernel methods," *J. Mach. Learn. Res.*, vol. 3, pp. 1439–1461, Mar. 2003.

[16] J. G. Dy and C. E. Brodley, "Feature selection for unsupervised learning," *J. Mach. Learn. Res.*, vol. 5, pp. 845–889, Jan. 2004.

[17] P. Mitra, C. A. Murthy, and S. K. Pal, "Unsupervised feature selection using feature similarity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 3, pp. 301–312, Mar. 2002.

[18] Z. Xu, I. King, M. R.-T. Lyu, and R. Jin, "Discriminative semi-supervised feature selection via manifold regularization," *IEEE Trans. Neural Netw.*, vol. 21, no. 7, pp. 1033–1047, Jul. 2010.

[19] Z. Zhao and H. Liu, "Semi-supervised feature selection via spectral analysis," in *Proc. SIAM Int. Conf. Data Mining*, Apr. 2007, pp. 641–646.

[20] B. Chandra and M. Gupta, "An efficient statistical feature selection approach for classification of gene expression data," *J. Biomed. Informat.*, vol. 44, no. 4, pp. 529–535, 2011.

[21] M. P. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Ares, Jr., and D. Haussler, "Knowledge-based analysis of microarray gene expression data by using support vector machines," *Proc. Nat. Acad. Sci. USA*, vol. 97, no. 1, pp. 262–267, Jan. 2000.

[22] K. Fujarewicz, M. Kimmel, and J. Rzeszowska-Wolny, "Improved classification of gene expression data using support vector machines," *J. Med. Informat. Technol.*, vol. 6, 2001.

[23] S. Mukherjee, P. Tamayo, D. Slonim, A. Verri, T. Golub, and J. Mesirov, *Support Vector Machine Classification of Microarray Data*. Cambridge MA, USA: Massachusetts Institute of Technology, 1999.

[24] A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, and Z. Yakhini, "Tissue classification with gene expression profiles," *J. Comput. Biol.*, vol. 7, nos. 3–4, pp. 559–583, 2000.

[25] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data," *Bioinformatics*, vol. 16, no. 10, pp. 906–914, 2000.

[26] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C.-H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J. P. Mesirov, T. Poggio, W. Gerald, M. Loda, E. S. Lander, and T. R. Golub, "Multiclass cancer diagnosis using tumor gene expression signatures," *Proc. Nat. Acad. Sci. USA*, vol. 98, no. 26, pp. 15149–15154, 2001.

[27] T. Li, C. Zhang, and M. Ogihara, "A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression ," *Bioinformatics*, vol. 20, no. 15, pp. 2429–2437, 2004.

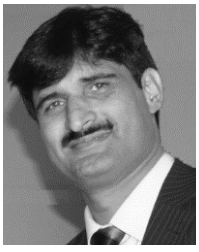
[28] E. Fix and J. L. Hodges, Jr, *Discriminatory Analysis-Nonparametric Discrimination: Consistency Properties*. Berkeley, CA, USA: California Univ. Berkeley, 1951.
[29] S. Dudoit, J. Fridlyand, and T. P. Speed, "Comparison of discrimination methods for the classification of tumors using gene expression data," *J. Amer. Statist. Assoc.*, vol. 97, no. 457, pp. 77–87, 2002.
[30] S. Zhang, X. Li, M. Zong, X. Zhu, and R. Wang, "Efficient kNN classification with different numbers of nearest neighbors," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 5, pp. 1774–1785, May 2018.
[31] Y. Lu, Q. Tian, M. Sanchez, J. Neary, F. Liu, and Y. Wang, "Learning microarray gene expression data by hybrid discriminant analysis," *IEEE Multimedia*, vol. 14, no. 4, pp. 22–31, Oct./Dec. 2007.
[32] A. Statnikov, L. Wang, and C. F. Aliferis, "A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification," *BMC Bioinf.*, vol. 9, no. 1, p. 319, 2008.
[33] J. R. Díaz-Uriarte, and S. A. de Andrés, "Gene selection and classification of microarray data using random forest," *BMC Bioinform.*, vol. 7, no. 1, p. 3, 2006.
[34] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.
[35] M. Z. Man, G. Dyson, K. Johnson, and B. Liao, "Evaluating methods for classifying expression data," *J. Biopharmaceutical Statist.*, vol. 14, no. 4, pp. 1065–1084, 2004.
[36] B. Wu, T. Abbott, D. Fishman, W. McMurray, G. Mor, K. Stone, D. Ward, K. Williams, and H. Zhao, "Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data," *Bioinformatics*, vol. 19, no. 13, pp. 1636–1643, 2003.
[37] J. Ma, Y. Qiao, G. Hu, Y. Huang, A. K. Sangaiah, C. Zhang, Y. Wang, and R. Zhang, "De-anonymizing social networks with random forest classifier," *IEEE Access*, vol. 6, pp. 10139–10150, 2017.
[38] M. Weedon, D. Tsaptsinos, and J. Denholm-Price, "Random forest explorations for URL classification," in *Proc. Int. Conf. Cyber Situational Awareness, Data Anal. Assessment (Cyber SA)*, Jun. 2017, pp. 1–4.
[39] R. Bernstein, M. Osadchy, D. Keren, and A. Schuster, "LDA classifier monitoring in distributed streaming systems," *J. Parallel Distrib. Comput.*, vol. 123, pp. 156–167, Jan. 2019.
[40] P. T. Noi and M. Kappas, "Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover classification using sentinel-2 imagery," *Sensors*, vol. 18, no. 1, p. 18, 2018.
[41] I. Ahmad, M. Basheri, M. J. Iqbal, and A. Raheem, "Performance comparison of support vector machine, random forest, and extreme learning machine for intrusion detection," *IEEE Access*, vol. 6, pp. 33789–33795, 2018.
[42] C. Chow, "On optimum recognition error and reject tradeoff," *IEEE Trans. Inf. Theory*, vol. IT-16, no. 1, pp. 41–46, Jan. 1970.
[43] B. Dubuisson and M. Masson, "A statistical decision rule with incomplete knowledge about classes," *Pattern Recognit.*, vol. 26, pp. 155–165, Jan. 1993.
[44] B. Hanczar and E. R. Dougherty, "Classification with reject option in gene expression data," *Bioinformatics*, vol. 24, no. 17, pp. 1889–1895, Sep. 2008.
[45] T. C. Landgrebe, D. M. Tax, P. Paclík, and R. P. W. Duin, "The interaction between classification and reject performance for distance-based reject-option classifiers," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 908–917, 2006.
[46] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, 1999.
[47] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proc. Nat. Acad. Sci. USA*, vol. 96, no. 12, pp. 6745–6750, 1999.
[48] M. J. Van De Vijver *et al.*, "A gene-expression signature as a predictor of survival in breast cancer," *New England J. Med.*, vol. 347, no. 25, pp. 1999–2009, 2002.
[49] J.-Y. Yeh, T.-S. Wu, M.-C. Wu, and D.-M. Chang, "Applying data mining techniques for cancer classification from gene expression data," in *Proc. Int. Conf. Converg. Inf. Technol. (ICCIT)*, Nov. 2007, pp. 703–708.
[50] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: Promise and potential," *Health Inf. Sci. Syst.*, vol. 2, no. 1, p. 3, 2014.
[51] E. Alpaydin, *Introduction to Machine Learning*, 3rd ed. Cambridge, MA, USA: MIT Press, 2014.

MUHAMMAD HAMMAD WASEEM is currently a Graduate Student with the Department of Computer Science and Information Technology, University of Azad Jammu and Kashmir.

MALIK SAJJAD AHMED NADEEM received the Ph.D. degree in Paris, France, in 2011. He is currently an Assistant Professor with the Department of Computer Sciences and Information Technology, University of Azad Jammu and Kashmir, Muzaffarabad. He has published various journal articles in the areas of machine learning, biomedical informatics, and decision support systems.

ASSAD ABBAS received the Ph.D. degree in electrical and computer engineering from North Dakota State University, USA. He is currently an Assistant Professor of computer science with COMSATS University Islamabad, Islamabad, Pakistan. His research has appeared in several reputable international venues. His research interests are mainly but not limited to smart health, big data analytics, recommendation systems, patent analysis, software engineering, and social network analysis. He is a member of the IEEE-HKN. He is also serving as the Referee for numerous prestigious journals and as a Technical Program Committee Member of several conferences.

ALIYA SHAHEEN is currently a Faculty Member with the Department of Mathematics, University of Azad Jammu and Kashmir, Pakistan.



WAJID AZIZ received the B.Sc. and M.Sc. degrees from the University of Azad Jammu and Kashmir (UAJ&K), Muzaffarabad, Pakistan, and the Ph.D. degree from the Pakistan Institute of Engineering and Applied Sciences (PIEAS), Islamabad, Pakistan. He started his career at UAJ&K, in 1998, as a Lecturer. He is currently a Professor with the College of Computer Science and Engineering, University of Jeddah, Jeddah, Saudi Arabia. He has published three books and more than 47 research articles in the reputed national and international journals and conference proceedings. His core research expertise is in biomedical information systems. His focused research areas include biomedical signal processing, time series analysis, and biomedical data analytics. Based on his academic and research contributions, he was a recipient of the HEC University Best Teacher Award for the year 2012–13 awarded by HEC Pakistan, in 2014, and the University Best Teacher Award by the University of AJ&K, in 2013.

ADEEL ANJUM received the Ph.D. degree (Hons.), in 2013. His area of research is data privacy using artificial intelligence techniques. He is currently an Assistant Professor with the Department of Computer Sciences, COMSATS University Islamabad (CIIT), Islamabad. He has several publications in reputed journals and international conferences. He is also the author of a book on data privacy. He serves on the technical program committees of various international conferences and journals.

UMAR MANZOOR received the B.S. and M.S. degrees in computer science from the National University of Computer and Emerging Sciences, Pakistan, in 2003 and 2005, respectively, and the Ph.D. degree in multiagent systems from the University of Salford, Manchester, U.K., in 2011. He is currently a Postdoctoral Researcher with the Computer Science Department, Tulane University, USA. He has strong interest in machine learning, natural language processing, multiagent systems, and artificial intelligence and has published extensively in these areas.

MUHAMMAD A. BALUBAID received the B.S. degree in manufacturing engineering and the M.S. degree in manufacturing engineering from the University of Warwick, U.K., and the Ph.D. degree in operations management and logistic from The University of Manchester, U.K. He is currently an Associate Professor with the Industrial Engineering Department, Faculty of Engineering, King Abdulaziz University, Jeddah, Saudi Arabia.

SEONG-O SHIM received the B.S. degree in electronics engineering from Ajou University, Suwon, South Korea, in 1999, and the M.S. degree in mechatronics and the Ph.D. degree in information and mechatronics from the Gwangju Institute of Science and Technology, Gwangju, South Korea, in 2001 and 2011, respectively. He was with LG Electronics DTV Labs, Seoul, South Korea, where he was involved in the research and development of digital TV, from 2003 to 2007. He is currently an Associate Professor with the College of Computer Science and Engineering, University of Jeddah, Jeddah, Saudi Arabia. His research interests include computer vision, image processing, 3D shape recovery, and medical imaging.