

[Click for updates](#)

## Applied Artificial Intelligence: An International Journal

Publication details, including instructions for authors and subscription information:

<http://www.tandfonline.com/loi/uaai20>

### Robust Feature Selection Technique Using Rank Aggregation

Chandrima Sarkar<sup>a</sup>, Sarah Cooley<sup>b</sup> & Jaideep Srivastava<sup>a</sup>

<sup>a</sup> College of Science and Engineering, University of Minnesota at Twin Cities, Twin Cities, Minnesota, USA

<sup>b</sup> Masonic Cancer Center, University of Minnesota at Twin Cities, Twin Cities, Minnesota, USA

Published online: 14 Mar 2014.

To cite this article: Chandrima Sarkar, Sarah Cooley & Jaideep Srivastava (2014) Robust Feature Selection Technique Using Rank Aggregation, Applied Artificial Intelligence: An International Journal, 28:3, 243-257, DOI: [10.1080/08839514.2014.883903](https://doi.org/10.1080/08839514.2014.883903)

To link to this article: <http://dx.doi.org/10.1080/08839514.2014.883903>

PLEASE SCROLL DOWN FOR ARTICLE

Taylor & Francis makes every effort to ensure the accuracy of all the information (the "Content") contained in the publications on our platform. However, Taylor & Francis, our agents, and our licensors make no representations or warranties whatsoever as to the accuracy, completeness, or suitability for any purpose of the Content. Any opinions and views expressed in this publication are the opinions and views of the authors, and are not the views of or endorsed by Taylor & Francis. The accuracy of the Content should not be relied upon and should be independently verified with primary sources of information. Taylor and Francis shall not be liable for any losses, actions, claims, proceedings, demands, costs, expenses, damages, and other liabilities whatsoever or howsoever caused arising directly or indirectly in connection with, in relation to or arising out of the use of the Content.

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden. Terms &



## ROBUST FEATURE SELECTION TECHNIQUE USING RANK AGGREGATION

**Chandrima Sarkar<sup>1</sup>, Sarah Cooley<sup>2</sup>, and Jaideep Srivastava<sup>1</sup>**

<sup>1</sup> College of Science and Engineering, University of Minnesota at Twin Cities,  
Twin Cities, Minnesota, USA

<sup>2</sup> Masonic Cancer Center, University of Minnesota at Twin Cities, Twin Cities,  
Minnesota, USA

□ Although feature selection is a well-developed research area, there is an ongoing need to develop methods to make classifiers more efficient. One important challenge is the lack of a universal feature selection technique that produces similar outcomes with all types of classifiers. This is because all feature selection techniques have individual statistical biases, whereas classifiers exploit different statistical properties of data for evaluation. In numerous situations, this can put researchers into dilemma with regard to which feature selection method and classifiers to choose from a vast range of choices. In this article, we propose a technique that aggregates the consensus properties of various feature selection methods in order to develop a more optimal solution. The ensemble nature of our technique makes it more robust across various classifiers. In other words, it is stable toward achieving similar and, ideally, higher classification accuracy across a wide variety of classifiers. We quantify this concept of robustness with a measure known as the robustness index (RI). We perform an extensive empirical evaluation of our technique on eight datasets with different dimensions, including arrhythmia, lung cancer, Madelon, mfeat-fourier, Internet ads, leukemia-3c, embryonal tumor, and a real-world dataset, vis., acute myeloid leukemia (AML). We demonstrate not only that our algorithm is more robust, but also that, compared with other techniques, our algorithm improves the classification accuracy by approximately 3–4% in a dataset with fewer than 500 features and by more than 5% in a dataset with more than 500 features, across a wide range of classifiers.

## INTRODUCTION

We live in an age of exploding information in which accumulating and storing data is easy and inexpensive. In 1991, it was pointed out that the amount of stored information doubles every twenty months (Piateski

Address correspondence to Chandrima Sarkar, College of Science and Engineering, University of Minnesota at Twin Cities, Department of Computer Science, 4-192 Keller Hall, 200 Union St., Minneapolis, MN, 55455 USA. E-mail: [sark0071@umn.edu](mailto:sark0071@umn.edu)

Color versions of one or more of the figures in the article can be found online at [www.tandfonline.com/uaai](http://www.tandfonline.com/uaai).

and Frawley 1991). Unfortunately, the ability to understand and utilize this information does not keep pace with its growth. Machine learning provides tools by which large quantities of data can be automatically analyzed. Feature selection is one of the fundamental steps of machine learning. Feature selection identifies the most salient features for learning and focuses a learning algorithm on those properties of the data that are most useful for analysis and future prediction. It has immense potential to enhance knowledge discovery by extracting useful information from high-dimensional data, as shown in previous studies in various important areas (Saeys, Inza, and Larrañaga 2007; Termenon et al. 2012; Wang and Song 2012; Yang and Pedersen 1997). In this article, we propose to develop an improved rank aggregation-based feature selection method that will produce a feature set that is more robust across a wide range of classifiers than the traditional feature selection techniques. In our previous work (Sarkar, Cooley, and Srivastava 2012) we developed the idea of a rank aggregation-based feature selection approach, and we showed that feature selection for supervised classification tasks can be accomplished on the basis of an ensemble of various statistical properties of data. In this study, we extend our idea by developing the rank aggregation-based feature selection algorithm with exclusive rank aggregation approaches such as Kemeny (1959) and Borda (1781). We evaluate our algorithm using five different classifiers over eight datasets with varying dimensions.

Feature selection techniques can be classified into filter and wrapper approaches (Guyon and Elisseeff 2003; Yu and Liu 2003). In this work, we focus on filter feature selection because it is faster and more scalable (Saeys, Inza, and Larrañaga 2007). Feature selection techniques using distinct statistical properties of data have some drawbacks; for example, information gain is biased toward choosing attributes with a large number of values, and chi square is sensitive to sample size. This indicates a statistical bias toward achieving the most optimal solution for the classification problem. In other words, there will be a variation in classification performance as a result of the partial ordering imposed by the evaluation measures (for example, information gain and chi square statistics) over the spaces of the hypotheses (Vilalta and Oblinger 2000). It has been shown that ensemble approaches reduce the risk of choosing a wrong hypotheses from many existing hypotheses in the solution space (Khoussainov, Hess, and Kushmerick 2005). Ensemble technique has also been used in various applications, showing notable improvement in the results, such as in Budnik and Krawczyk (2013), Woźnial, Graña, and Corchado (2014), and Abeel and colleagues (2010). To the best of our knowledge, no other study has focused on an extensive performance evaluation of rank aggregation-based feature selection technique using exclusive rank aggregation strategies such as Kemeny (1959).

To summarize, this article makes the following contributions:

1. We develop a novel rank aggregation-based feature selection technique using exclusive rank aggregation strategies by Borda (1781) and Kemeny (1959).
2. We make an extensive performance evaluation of the rank aggregation-based feature selection method using five different classification algorithms over eight datasets of varying dimensions. We perform pairwise statistical tests with 5% significance level to prove the statistical significance of the classification accuracy results.
3. We introduce the concept of robustness of a feature selection algorithm, which can be defined as the property that characterizes the stability of a ranked feature set toward achieving similar classification accuracy across a wide range of classifiers. We quantify this concept with an evaluation measure named the robustness index (RI).
4. We propose to use the concept of inter-rater agreement for improving the quality of the rank aggregation approach for feature selection.

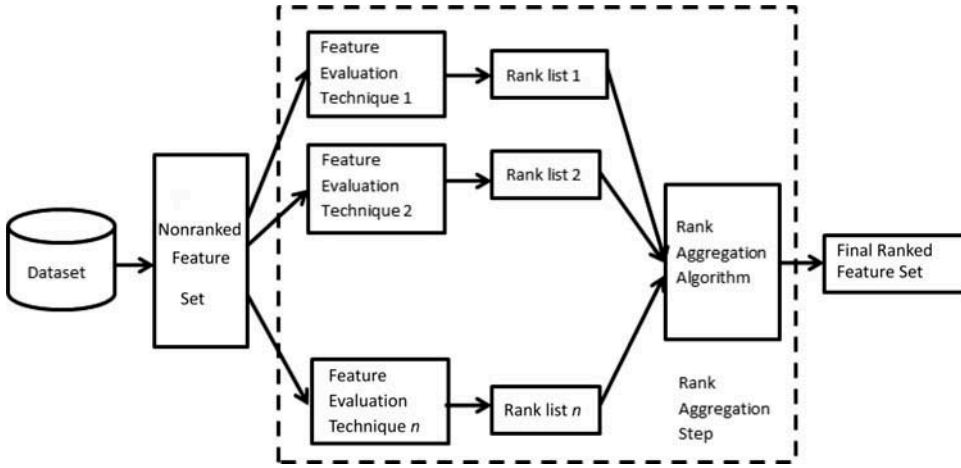
The remainder of this article is organized as follows. The following section presents the “Methodology”; the section following that describes the “Experimental Results and Discussion.” In the final section, we present our conclusion. In this article, we use the terms *variables*, *features*, and *attributes* interchangeably.

## METHODOLOGY

The process of rank aggregation-based feature selection technique consists of the following steps: a nonranked feature set is evaluated with  $n$  feature selection/evaluation techniques. This gives rise to  $n$  sets of ranked feature sets, which differ in their rank ordering. The following step consists of executing rank aggregation on the feature sets using either Borda (1781) or Kemeny–Young (Kemeny 1959) strategy to generate a final ranked feature set. The entire process of rank aggregation is documented inside the dotted box in Figure 1.

### Rank Aggregation

Rank aggregation is the process that combines the ranking results of a fixed set of candidates from multiple ranking functions to generate a single better ranking. Rank aggregation can be done in different ways in other words, the Borda method (Borda 1781) and the Kemeny method (Kemeny 1959). The rank aggregation step has been described in Algorithm 1.



**FIGURE 1** Flow diagram of the rank aggregation-based feature selection algorithm. The number of feature evaluation techniques (using different statistical properties of data) is represented by  $n$ .

---

**Algorithm 1.** Rank Aggregation

---

**Input:**

1) Feature Evaluation Technique (FET) set  $Q'$  and feature set  $S$

**Output:**

Ranked Feature Set  $S'$

**Algorithm:**

**Steps:** Rank Aggregation

**2: for**  $j = 0$  to  $|Q'| - 1$  **do**

$S'' = S$

$\triangleright S''$  is a temporary variable

4: Rank  $S''$  using  $FET_j$  where  $FET_j \in Q'$   
add  $S''$  to list  $L$

**6: end for**

$S' = \text{aggregated feature set from } L \text{ using Equation (1)}$

---

### Borda Method

In this study, we use rank aggregation based on Borda (1781) ranking. For this we use a position-based scoring mechanism to calculate the score of a feature. A predetermined score is dedicated to each position in a list generated from each feature selection technique (this score is the same for all the lists). For a distinct feature, the final score is the sum of all the positional scores from all the lists, as given in Equation (1). The final rank of a feature is determined from the final score:

$$\text{score}_{\text{final}} = \sum_{i=1}^n \text{score}_{\text{pos}(i,j)}, \quad (1)$$

where  $n$  is the total number of feature selection techniques (or rankers) used,  $pos(i, j)$  is the  $j$ th position of a feature ranked by the ranker  $i$ ,  $score_{p(i,j)}$  is the score of a feature in list  $i$  generated by ranker  $i$  at the  $j$ th position, and  $score_{final}$  is the sum of all the positional scores from all the lists. In this work, we consider a single-feature selection technique as a ranker and the candidates as the features.

### **Kemeny Method**

The Kemeny rule is sometimes interpreted as a maximum likelihood estimator of the “correct” ranking (Conitzer, 2006), and for every pair of candidates, any voter ranks the better candidate higher, with probability  $p > \frac{1}{2}$ , independently. The Kemeny rule is given as follows: Let  $x_1$  and  $x_2$  be two candidates,  $r$  be a ranking and  $v$  be the vote, let  $\delta_{x_1, x_2}(r, v) = 1$  if there exists an agreement between  $r$  and  $v$  on the relative ranking of  $x_1$  and  $x_2$  that is either both rank  $x_1$  as higher, or both rank  $x_2$  as higher, or else 0 if they disagree. Let  $T'$  be the total number of pairwise agreements, that is, the agreement of a ranking  $r$  with a vote  $v$ , which is given by  $\sum_{x_1, x_2} \delta_{x_1, x_2}(r, v)$ . Then a Kemeny ranking  $r$  maximizes the sum of the agreements with the votes given by  $\sum_v \sum_{x_1, x_2} \delta_{x_1, x_2}(r, v)$  (Conitzer 2006). Because computing optimal Kemeny aggregation is NP-hard for  $r \geq 4$  (Dwork et al. 2001), in this study we use the 2-approximation of Kemeny optimal aggregation (Subbian and Melville 2011), which has been shown to run in time  $O(R * c \log c)$ , where  $R$  denotes the total number of rankers and  $c$  denotes the number of candidates.

### **Analysis of Rankers with Inter-Rater Agreement**

Inter-rater agreement (IRA) can be used as a preprocessing step prior to the rank aggregation step. The main motivation for this step is the analysis of the homogeneity or consensus among the rank ordering generated by each ranker. Each ranker uses different measures for evaluating the candidates and, hence, generates a different rank ordering. There can be a possibility that the rank ordering generated by one of the rankers is highly inconsistent with the other rankers. This might cause the final aggregated rank ordering to be far away from the ground truth (optimal) ordering. In this article, we make the assumption that rank ordering that is in consensus with the majority of the rankers is closest to the ground truth. Hence, for improving the rank ordering generated by the ranker, we propose to use the concept of IRA, which analyzes the degree of agreement among rankers. We use the intraclass correlation (ICC; Bartko 1976) approach for calculating the IRA. The ICC assesses rating reliability by comparing the variability of different

ratings of the same subject to the total variation across all ratings and all subjects, as formulated in Equation (2):

$$ICC = \frac{V(b)^2}{(V(w)^2 + V(b)^2)}, \quad (2)$$

where  $V(w)^2$  is the pooled variance within subjects, and  $V(b)^2$  is the variance of the trait between subjects. The IRA lies between 0 and 1, where 0 is the least-reliable rating and 1 is the most-reliable rating for a group of rankers. We use a heuristically determined threshold  $T$ , which is used for eliminating rankers who tend to disrupt the homogeneity in ranking from the group of rankers.

### ***K-step Feature Subset Selection***

The  $K$ -step feature subset selection is a postprocessing step to the rank aggregation step, with a focus on generating a feature subset from the final rank aggregated feature set. In this process, first, for each classification algorithm, we determine the classification accuracy of each top  $i$  feature subset  $1 \leq i \leq k$ , where  $k$  is the total number of features in the feature subset. Next, we select the feature subset with the maximum classification accuracy across all the classification algorithms used as our final feature subset, as given in Algorithm 2.

---

#### **Algorithm 2.** $K$ -step Feature Subset Selection

---

**Input:**

Feature Set  $S'$ , Dataset  $D$ , Classifiers set  $M$

**Output:**

feature subset of size  $k$

**Algorithm:**

set  $S^* \subset S'$ ,  $m_i \in M$  and  $\sum_i |m_i| = |M|$  and  $K_j \in S'$  with  $K_1 < K_2 \dots K_j \dots < K_k$  where  $K$  is the feature subset,  $1 > j > k$

2: **for**  $1 < i < |M|$  **do**

**for**  $1 < j < K$  **do**

4:     add feature set  $K_j$  to  $S^*$   
       learn  $S^*$  using  $M_i$

6:     Calculate accuracy of  $M_i$  and store it in list  $temp_j$

**end for**

8:   search  $temp_j$  for  $1 < j < K$  with the highest predictive power for  $M_i$  and store in  $K*_i$

**end for**

10: select the  $\text{MAX}(K*_i)$  where  $1 < i < |M|$

---

### ***Evaluation Measures***

We evaluate our algorithm based on three evaluation measures as discussed following:



1. Classification accuracy. Accuracy is calculated as the percentage of correctly classified instances by a given classifier. At first we obtain a feature subset of size  $K$  using a  $K$ -step feature subset selection approach as described in the previous section. Classification accuracy of this feature subset was determined and recorded using five different classifiers for evaluation purposes.
2. F-measure. Weighted (by class size) average F-measure was obtained from the classification using a feature subset with the same five classifiers as in evaluation measure 1.
3. Robustness Index (RI) - We define robustness as the property that characterizes the stability of a feature subset toward achieving similar classification accuracy across a wide range of classifiers. In order to quantify this concept, we introduce a measure called the robustness index (RI), which can be utilized for evaluating the robustness of a feature selection algorithm across a variety of classifiers. Intuitively, RI measures the consistency with which a feature subset generates similar (ideally higher) classification accuracy (or lower classification error) across a variety of classification algorithms when compared with feature subsets generated by using other methods. The step-by-step process of the robustness index is described in Algorithm 3.

---

**Algorithm 3.** Robustness Index Calculation

---

**Input**

- 1) Classification models  $M_i$ , where  $1 \leq i \leq m$ ,  $m$  is the number of classifiers used,
- 2) Feature set  $f_k$  generated from  $p$  feature selection techniques, where  $k$  is number of top features

**Output**

Robustness Index  $r_p$ , for each  $p$  feature selection technique

**Algorithm:**

```

for  $i = 0$  to  $m - 1$  do                                     ▷ For each  $M_i$ 
2:   for  $j = 0$  to  $p - 1$  do
        $C_p$  = classification error with  $f_k$ 
4:   end for
       Rank each  $p$  based on  $C_p$  score and save
6: end for
   for  $i = 0$  to  $p - 1$  do
8:   Aggregate the ranks across  $M_i$  for  $1 < i < m$  using Equation (1)
       Assign  $r_p$  = aggregated ranks
10: end for

```

---

The motivation behind the concept of robustness is as follows: It is not an easy task to determine the best classifier to use for a classification task prior to actually using that model. A robustness technique helps one to choose a classification model with the minimum risk in choosing an inappropriate model.

TABLE 1 Data Sets with Attributes and Instances

Datasets		
Dataset	Attributes	Instances
Lung Cancer	57	32
AML	69	927
Mfeat-fourier	77	2000
Arrhythmia	280	452
Madelon	501	2600
Internet-Ads	1559	3279
Leukemia-3c	7130	72
Embryonal Tumor	7130	60

Experimental Setup

Datasets Used

We used eight different types of datasets as shown in Table 1. Acute myeloid leukemia, or AML, is a real-world dataset that contains 69 demographic, genetic, and clinical variables from 927 patients who received myeloablative, T-cell replete, unrelated donor (URD) stem cell transplants (Sarkar, Cooley, and Srivastava 2012). We also used datasets including embryonal tumors of the central nervous system (Pomeroy et al. 2002), Madelon and Internet-ads (Pathical 2010), leukemia-3c, arrhythmia, lung cancer and mfeat-fourier (Hettich and Bay 1999) from UCI KDD as listed in Table 1.

Statistical Test of Significance

We performed a pairwise *t*-test with a 5% significance level in order to measure the statistical significance of our classification accuracy result. The null hypothesis is that the difference between the classification accuracy results obtained from the two algorithms considered in the pairwise test comes from a normal distribution with mean equal to zero and unknown variance. We reject the null hypothesis if the *p*-value is less than 5% significance level. The result is given in Table 2.

EXPERIMENTAL RESULTS AND DISCUSSION

We evaluated the performance of our feature selection algorithm in terms of classification accuracy, F-measure, and robustness by comparing with three feature selection techniques: information gain attribute evaluation, symmetric uncertainty attribute evaluation, and chi square attribute evaluation (Hall et al. 2009). We selected feature selection techniques with

TABLE 2 Results of Paired *t*-test for Different Datasets

Lung Cancer			AML		mfeat - fourier		Embryonal Tumo	
Borda	Kemeny		Borda	Kemeny	Borda	Kemeny	Borda	Kemeny
IG	0.0327	0.0401	IG	0.0509	0.0212	0.0227	IG	0.0155
SU	0.009	0.0074	SU	0.0571	0.0227	0.0243	SU	2.7E-04
CS	0.0046	0.0066	CS	0.0509	0.0227	0.0243	CS	0.0225
Madelon			Internet-Ads		Leukemia-3c		Arrhythmia	
Borda	Kemeny		Borda	Kemeny	Borda	Kemeny	Borda	Kemeny
IG	0.0164	0.0164	IG	0.0119	0.0015	0.0015	IG	0.0157
SU	0.0139	0.0139	SU	0.0119	5.3E-04	5.3E-04	SU	0.0049
CS	0.0164	0.0164	CS	0.0119	3.8E-04	3.8E-04	CS	0.0034

IG - Information gain, SU - Symmetric Uncertainty, CS - Chi Square.

the help of the IRA method, assuming an IRA threshold of 0.75 (heuristically determined). We refer our rank aggregation-based feature selection algorithm as *Kemeny* and *Borda* (using *Kemeny* and *Borda* methods, respectively) in the figures shown in this article.

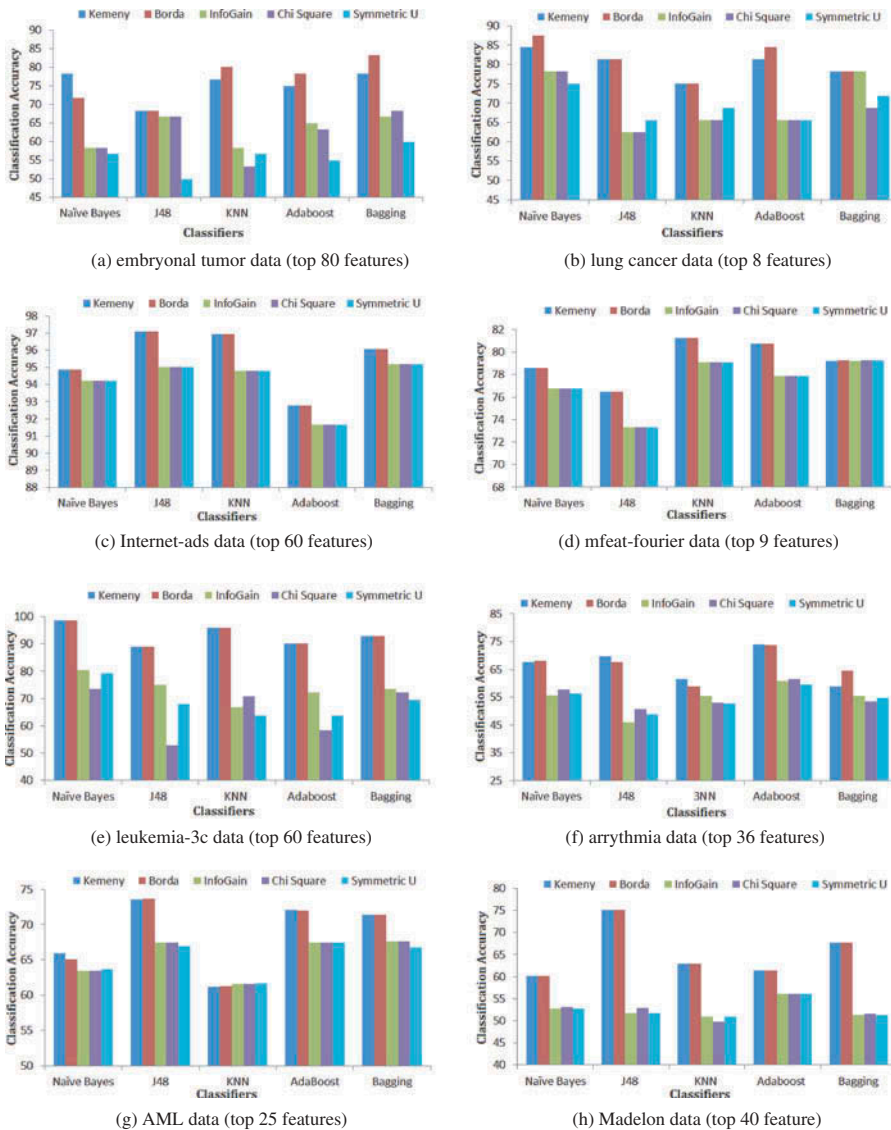
The results of classification accuracy over eight datasets are given in Figure 2. Using Algorithms 1 and 2, we generate feature subsets for each dataset indicated in a bracket in the Tables 3 and 5. Next, we perform classification using five different classifiers (shown in Table 4) whose results are given in Figures 2a, 2b, 2c, 2d, 2e, 2f, 2g, 2h. Figures 2a, 2c, 2h, and 2e show classification accuracy for datasets with over 500 variables. In these four datasets, the accuracy with *Kemeny* and *Borda* is more than 5% higher compared to those with the three single-feature selection methods. In the other four datasets shown in Figures 2b, 2d, 2f and 2g, the classification accuracy is higher by approximately 3–4 % across all the classifiers.

Next, we performed the pairwise statistical significance test with a 5% significance level to prove the statistical significance of our accuracy results. We calculated the  $p$ -values for every dataset, comparing *Kemeny* and *Borda* with three feature selection techniques, as depicted in Table 2.

Table 5 shows the comparison of the robustness index as calculated using Algorithm 3. The lower the value of RI, the more robust is the technique; in other words, the RI that equals 1 is more robust than an RI that equals 3. We can see from Table 5 that both *Kemeny* and *Borda* have RIs of either 1 or 2 with every dataset. This shows that *Kemeny* and *Borda* are more robust than the other traditional feature selection techniques. The motivation behind this analysis is that, when one is unable to decide on the best classification algorithm to use on a given data set, our feature selection algorithm will help with a technique that will ensure a lower classification error over a variety of classification algorithms. The numbers in parentheses beside the dataset names in Table 5 indicates the size of the feature subset.

In Table 3, we compare the weighted (by class size) average F-measures (defined as the harmonic means of precision and recall), generated using *Kemeny* and *Borda* with three feature selection methods using five different classifiers, as given in Table 4. The numbers in parentheses beside the dataset names in every figure indicate the size of the feature subset used for classification. We can see that F-measure with *Kemeny* and *Borda* is higher in almost all the cases. This shows that, apart from accuracy, the sensitivity and specificity generated with different classifiers using our rank aggregation-based feature selection method can be improved.

The results of these analyse show that our rank aggregation-based feature selection algorithm is an efficient technique suited for various kinds of datasets including those with features greater than 1000. Our method gives a higher classification accuracy, F-measure, and greater robustness than the



**FIGURE 2** Comparison of classification accuracies in different datasets using *Kemeny*, *Borda*, and three single-feature selection techniques.

other traditional methods over a wide range of classifiers. This method is advantageous especially in cases for which it is difficult to determine the best statistical property for evaluation of a given dataset. The greatest advantage in having a robust technique is that there will be fewer dilemmas when deciding on the most appropriate classifier to use from the vast range of choices.

TABLE 3 Results for F-Measure

	Naïve Bayes Classifier				
	Kemeny	Borda	IG	SU	CS
Datasets					
AML (25)	0.67	0.67	0.64	0.65	0.64
Mfeat-fourier (9)	0.79	0.79	0.77	0.77	0.77
Arrhythmia (36)	0.67	0.67	0.53	0.56	0.54
Madelon (40)	0.60	0.60	0.53	0.53	0.53
Internet-Ads (60)	0.95	0.95	0.94	0.94	0.94
Leukemia-3c (60)	0.99	0.99	0.81	0.73	0.79
Embryonal Tumor (60)	0.74	0.72	0.62	0.59	0.61
Lung Cancer (8)	0.45	0.45	0.46	0.51	0.40
J48 Decision Tree					
	Kemeny	Borda	IG	SU	CS
Datasets					
AML (25)	0.69	0.69	0.64	0.64	0.64
Mfeat-fourier (9)	0.76	0.76	0.73	0.73	0.73
Arrhythmia (36)	0.65	0.65	0.46	0.50	0.48
Madelon (40)	0.75	0.75	0.51	0.53	0.51
Internet-Ads (60)	0.97	0.97	0.95	0.95	0.95
Leukemia-3c (60)	0.89	0.89	0.75	0.51	0.68
Embryonal Tumor (60)	0.68	0.68	0.73	0.72	0.73
Lung Cancer (8)	0.48	0.48	0.47	0.43	0.42
k-Nearest Neighbor					
	Kemeny	Borda	IG	SU	CS
Datasets					
AML (25)	0.63	0.63	0.60	0.62	0.60
Mfeat-fourier (9)	0.83	0.83	0.81	0.81	0.81
Arrhythmia (36)	0.62	0.60	0.50	0.48	0.48
Madelon (40)	0.67	0.67	0.50	0.51	0.50
Internet-Ads (60)	0.97	0.97	0.95	0.95	0.95
Leukemia-3c (60)	0.96	0.96	0.63	0.68	0.61
Embryonal Tumor (60)	0.77	0.79	0.64	0.63	0.64
Lung Cancer (8)	0.38	0.38	0.41	0.28	0.52
AdaBoost					
	Kemeny	Borda	IG	SU	CS
Datasets					
AML (25)	0.69	0.69	0.63	0.63	0.63
Mfeat-fourier (9)	0.83	0.83	0.81	0.81	0.81
Arrhythmia (36)	0.45	0.45	0.42	0.42	0.42
Madelon (40)	0.61	0.61	0.54	0.54	0.54
Internet-Ads (60)	0.92	0.92	0.91	0.91	0.91
Leukemia-3c (60)	0.90	0.90	0.71	0.59	0.63
Embryonal Tumor (60)	0.73	0.73	0.57	0.56	0.57
Lung Cancer (8)	0.47	0.47	0.47	0.48	0.47

(Continued)

TABLE 3 (Continued)

	Bagging				
	Kemeny	Borda	IG	SU	CS
Datasets					
AML (25)	0.69	0.69	0.63	0.64	0.63
Mfeat-fourier (9)	0.79	0.79	0.78	0.78	0.78
Arrhythmia (36)	0.71	0.71	0.54	0.55	0.53
Madelon (40)	0.79	0.79	0.54	0.54	0.54
Internet-Ads (60)	0.96	0.96	0.95	0.95	0.95
Leukemia-3c (60)	0.93	0.93	0.73	0.70	0.69
Embryonal Tumor (60)	0.76	0.76	0.62	0.57	0.59
Lung Cancer (8)	0.44	0.44	0.35	0.35	0.33

TABLE 4 Classification Algorithms Used

Classifiers		Settings
Naive Bayes	estimator classes	
J48	pruned C4.5 decision tree	
KNN	k=3; brute force search algorithm; Euclidean distance function	
AdaboostM1	base classifier: Decision Stump,10 boost iterations, % of weight mass for training was 100	
Bagging	weak classifier: fast decision tree learner, bag size as 100%, 10 bagging iterations	

TABLE 5 Robustness Index Comparison

	Feature Selection Techniques				
	Kemeny	Borda	InfoGain	SymmetricU	Chi Square
Datasets					
AML (25)	1	1	3	2	3
Lung Cancer (8)	2	1	3	3	4
Arrhythmia (36)	1	2	4	5	3
mfeat-fourier (9)	2	1	4	3	3
Madelon (40)	1	1	2	3	2
Internet-Ads (60)	1	1	2	2	2
Leukemia-3c (60)	1	1	2	3	3
Embyonal Tumor (60)	2	1	3	4	5

CONCLUSION

In this article, we propose a novel rank aggregation-based feature selection technique beneficial for classification tasks. The results of our algorithm suggest that our proposed ensemble technique yields higher classification accuracy, higher F-measure, and greater robustness than single-feature selection techniques on datasets with different ranges of dimensions. The eight different datasets that we used have dimensions from as low as 57 (lung

cancer dataset) to as high as 7130 (leukemia-3c dataset). We found that our algorithm improved accuracy, F-measure, and robustness of classification in all the datasets. We proved the statistical significance of our classification accuracy results using a pairwise *t*-test with a 5% significance level. This shows that our feature selection technique is suited for high-dimensional data applications, especially in situations for which it is difficult to determine the best statistical property to use for evaluation of a feature set. We conclude by stating that our robust feature selection technique is an appropriate approach to be utilized in situations when one faces the dilemma of choosing the most suitable classifiers and the best statistical property to use for an improved result on a given dataset. Our experiments and the results provide initial evidence for the success of our feature selection framework. We believe that this framework has the potential to improve the accuracy and robustness of various classification tasks in many different applications.

## FUNDING

AML data resource in this work was supported by the National Institutes of Health/NCI grant P01 111412, PI Jeffrey S. Miller, M.D, utilizing the Masonic Cancer Center, University of Minnesota Oncology Medical Informatics shared resources. We would like to thank Atanu Roy for his critical reviews and technical feedback during the development of this research.

## REFERENCES

- Abeel, T., T. Helleputte, Y. Van de Peer, P. Dupont, and Y. Saeys. 2010. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics* 26(3):392–398.
- Bartko, J. 1976. On various intraclass correlation reliability coefficients. *Psychological Bulletin* 83(5):762.
- Budnik, M., and B. Krawczyk. 2013. On optimal settings of classification tree ensembles for medical decision support. *Health Informatics Journal* 19(1):3–15.
- Conitzer, V. 2006. *Computational aspects of preference aggregation*. PhD thesis, IBM.
- Borda, J. C. de. 1781. *Memoire sur les Elections au Scrutin*. Histoire de l'Academie Royale des Sciences, Paris.
- Dwork, C., R. Kumar, M. Naor, and D. Sivakumar. 2001. Rank aggregation methods for the web. In *Proceedings of the 10th international conference on world wide web*, 613–622. ACM.
- Guyon, I., and A. Elisseeff. 2003. An introduction to variable and feature selection. *The Journal of Machine Learning Research* 3: 1157–1182.
- Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten. The weka data mining software: An update. *ACM SIGKDD Explorations Newsletter* 11(1):10–18.
- Hettich, S., and S. Bay. 1999. The uci kdd archive.
- Kemeny, J. 1959. Mathematics without numbers. *Daedalus* 88(4):577–591.
- Khoussainov, R., A. Hess, and N. Kushmerick. 2005. Ensembles of biased classifiers. In *Proceedings of the 22nd international conference on machine learning*, 425–432. ACM.
- Pathical, S. P. 2010. *Classification in high dimensional feature spaces through random subspace ensembles*. PhD thesis, University of Toledo.
- Piatetski, G., and W. Frawley. 1991. *Knowledge discovery in databases*. Cambridge, MA, USA: MIT press.
- Pomeroy, S. L., P. Tamayo, M. Gaasenbeek, L. M. Sturla, M. Angelo, M. E. McLaughlin, J. Y. Kim, L. C. Goumnerova, P. M. Black, C. Lau, et al. 2002. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature* 415(6870):436–442.
- Saeys, Y., I. Inza, and P. Larrañaga. 2007. A review of feature selection techniques in bioinformatics. *Bioinformatics* 23(19):2507–2517.



- Sarkar, C., S. Cooley, and J. Srivastava. 2012. Improved feature selection for hematopoietic cell transplantation outcome prediction using rank aggregation. In *Federated Conference on Computer Science and Information Systems (FedCSIS)*, 2012, 221–226. IEEE.
- Subbian, K., and P. Melville. 2011. Supervised rank aggregation for predicting influence in networks. *arXiv preprint arXiv:1108.4801*.
- Termenon, M., M. Graña, A. Besga, J. Echeveste, and A. Gonzalez-Pinto. 2012. Lattice independent component analysis feature selection on diffusion weighted imaging for Alzheimer's disease classification. *Neurocomputing* 114:132–141.
- Vilalta, R., and D. Oblinger. 2000. A quantification of distance bias between evaluation metrics in classification. In *Proceedings of the international conference on machine learning*, 1087–1094. Stanford University, Stanford, CA, June.
- Wang, G., and Q. Song. 2012. Selecting feature subset for high dimensional data via the propositional foil rules. *Pattern Recognition* 45(7):2672–2689.
- Woźniak, M., M. Graña, and E. Corchado. 2014. A survey of multiple classifier systems as hybrid systems. *Information Fusion* 16:3–17.
- Yang, Y., and J. Pedersen. 1997. A comparative study on feature selection in text categorization. In *Machine learning-international workshop then conference*, 412–420. Morgan Kaufmann.
- Yu, L., and H. Liu. 2003. Feature selection for high-dimensional data: A fast correlation-based filter solution. In *Machine learning-international workshop then conference* 20:856.