I.J. Information Technology and Computer Science, **2025, 2, 49-59**
Published Online on April 8, 2025 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2025.02.05

RSKD Ensemble Classifier with Stable Ensemble Feature Selection for High Dimensional Low Sample Size Cancer Datasets

Archana Suhas Vaidya*
GES's R H Sapat College of Engineering Management Studies and Research, Nashik 422005 Affiliated to Savitribai Phule Pune University, Pune, Maharashtra, India
E-mail: archana.vaidya@ges-coengg.org
ORCID iD: https://orcid.org/0000-0002-6213-6216
*Corresponding author

Dipak V. Patil
GES's R H Sapat College of Engineering Management Studies and Research, Nashik 422005 Affiliated to Savitribai Phule Pune University, Pune, Maharashtra, India
E-mail: dipakvpatil17@gmail.com
ORCID iD: https://orcid.org/0000-0002-9612-8564

Received: 05 October 2024; Revised: 17 December 2024; Accepted: 05 February 2025; Published: 08 April 2025

Abstract: This study presents the RSKD ensemble classifier, developed with ensemble feature selection techniques, to address high-dimensional, low-sample-size cancer datasets. Ensemble classifiers are advantageous in such scenarios, offering better classification accuracy than traditional methods by combining multiple models. This combination enhances predictive performance on high-dimensional datasets. However, stability—a key factor for consistent performance on unseen data—often involves a tradeoff with accuracy. Ensemble methods, due to their generalization capabilities, exhibit higher stability, with feature selection stability measured using a consistency index, averaging 65–70%.

The RSKD classifier integrates ensemble feature selection methods SU-R and ChS-R, which enhance feature selection stability and classification accuracy. Its performance was evaluated on seven high-dimensional, low-sample-size datasets and compared against state-of-the-art classifiers, including Adaboost, GradientBoost, REPTree, asBagging_FSS, SRKNN, MF-GE, and eAdaBoost with DSC. The RSKD ensemble classifier achieved an accuracy improvement of 7.69% to 12.35% over these methods. Among the feature selection approaches, SU-R combined with RSKD outperformed ChS-R, demonstrating superior results in cancer prediction tasks.

The findings of this study underscore the potential of RSKD for achieving generalized, robust performance on challenging datasets. By leveraging ensemble classifiers and ensemble feature selection techniques, researchers can address the inherent difficulties of high-dimensional, low-sample-size datasets, enhancing both accuracy and stability. This work provides a valuable foundation for developing diverse, heterogeneous ensemble approaches for cancer prediction and similar applications.

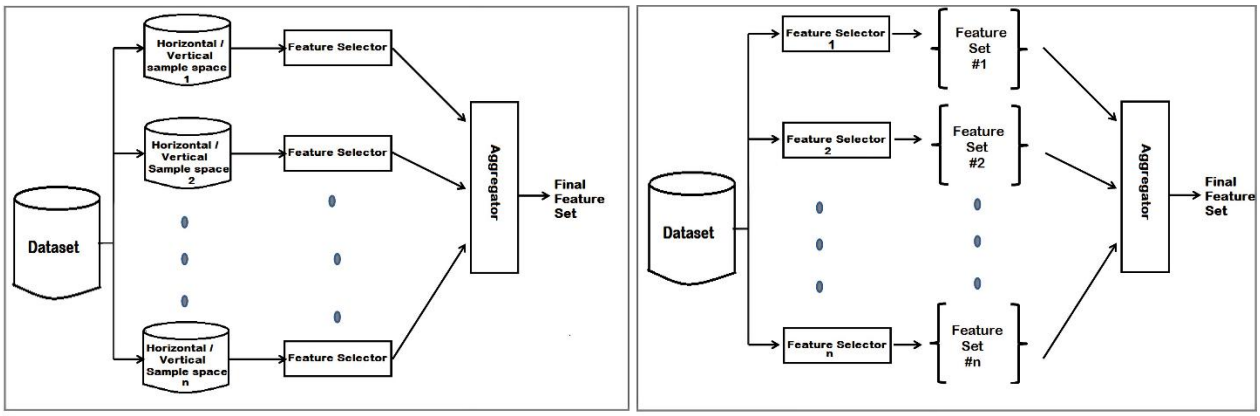
Index Terms: High Dimensional, Low Sample Dataset, Ensemble Feature Selection, Stability Analysis, Ensemble Classifier, Heterogeneous Ensemble.

1. Introduction

An ensemble methods for data mining and pattern discovery are growing and current research topic. These methods are being used more frequently to process the massive amounts of high-dimensional data produced by the high-throughput technologies. Training classifiers on skewed data, especially when the data is highly dimensional is an important problem. Biomedical researchers frequently use skewed data. The ensemble methods, such as bagging and boosting train several learners before combining them. In every ensemble strategy, combining incomplete results to produce a final result is a crucial step. It is well acknowledged that a group of people working together to achieve a common goal is usually much more accurate than a single person. The group learning techniques have been shown to be very successful in a range of real-world applications [1].

The various levels that can be leveraged to construct various forms of ensembles for feature selection, such as incorporating diverse combination methods, ensemble feature ranking, different base learners, dataset partitioning and searching ensembles. Various ensemble ranking approaches can be used for feature reduction in the initial phases. It is possible to use ensemble searching in a reduced feature set. The classification of ensemble can be done as homogeneous and Heterogeneous.

The ensemble is homogeneous if all of the base selectors are of the same type; otherwise, the ensemble is heterogeneous. The same feature selection procedure is employed in the homogeneous approach as shown in Figure 1a,with various training data subsets or feature subsets. These data subsets can be spread across multiple nodes (or partitions), resulting in a reduction in time requirements. The size of the partitions is also a design component in this method. A data variation ensemble is another term for these methods. Some examples of homogeneous approaches capable of managing large scale scenarios, can be found in [1,2] and to deal with imbalanced data sets [3] are more popular than homogeneous approaches. As shown in Figure 1 b), the heterogeneous approach applies a variety of different feature selection algorithms to the same training data. As previously indicated, the number of different feature selectors to be used in this approach is a design parameter. These techniques are sometimes known as function variation techniques. The heterogeneous approaches are more popular than homogeneous approaches. Several examples can be found in [4-8]. Mixed approaches can be employed with data variation and learner variation and can be found in [1, 9-11]



a) Homogeneous ensemble for feature selection b) Heterogeneous ensemble for feature selection

Fig.1. Different working models of ensemble feature selection

The rest of this paper is organized as follow. The current state of the art in ensemble classification and ensemble feature selection are discussed in Sect. 2. The proposed model assessment for stability analysis of our ensemble techniques for high dimensional low sample size cancer datasets is in Sect. 3. The accuracy testing of our proposed ensemble classifier and discussions of the results presented in Sects. 4. Sect. 5 presents the conclusions and future work.

2. Related Works

There are two primary methods for categorizing new instances in ensemble classification. During the classification phase of the first approach, the categories are fused in various ways. In the second technique, one classifier's categorization is chosen based on a criterion.

Fusing methods combine the results of multiple classifiers to provide classification. Weighting methods and metal earning methods are two subcategories of fusion methods.

2.1. Voting Methods

Hard Majority Voting is a majority voting where class label is mode of majority classifier class label and calculated as following eq. (1).

$$\hat{y} = \text{mode}\{C_1(x), C_2(x), \ldots, C_m(x)\} \quad (1)$$

Example suppose there are three classifiers
$C_1 \rightarrow$ class 0
$C_2 \rightarrow$ class 1
$C_3 \rightarrow$ class 0

Then final class label will be class 0 as here majority classifier vote is class 0.

Soft majority voting is calculated as per eq. (2). Where w_i is weight given to jth classifier. Final class label is

calculated by weighted majority classifier label.

$$\hat{y} = \arg\max \sum_{j=1}^{m} w_i \; p_{ij} \tag{2}$$

Performance Voting uses a performance evaluation such as accuracy of classifier i on a validation set. The class with the greatest score is selected as a final label. Various performance metrics such as accuracy, misclassification rate are discussed in [12].

Distribution Summation works on the conditional probability [13] vectors produced from each classifier. The selected class is determined by the whole vector's highest value. It can be expressed mathematically as:

$$Class(x) = \underset{Ci \in dom(y)}{\arg max} \sum_k \hat{P}M_k \; (y = C_i|x) \tag{3}$$

2.2. Decomposition Techniques

When homogeneous selectors as a ensemble are applied for feature selection data decomposition is done to divide the data. There are various strategies of decomposition as horizontal, vertical and mixed is described in Table 1. Parallel computing can be easily adopted to decrease the time complexity of ensemble feature selection with various decomposition techniques. Along with data partitioning various base learners can be applied in model selection process are studied in [14].

Table 1. Decomposition Techniques horizontal, vertical and Mixed /Hybrid ensemble feature selection methods working principle

Method	Description	Ref.
Horizontal or Tuple Based Partitioning	Sampling is used to divide the training tuples into subgroups. Sampling, a degenerate kind of decomposition that reduces complexity, falls within this group. Tuple based decomposition can be done in parallel, allowing for parallel learning.	[15-20]
Vertical or Feature based Partitioning	As per the values of the selection of input feature, the training tuples are separated into subsets. This is also termed as “feature value decomposition”[26]	[21-27]
Mixed or Hybrid Partitioning	Horizontal and vertical partitioning of data is mixed for ensemble feature selection. First data is divided horizontally and then vertically in individual partition. Parallel processing can be done to reduce time complexity.	[28-30]

Table 2. Ensemble classifier methods compared based on base classifiers used and feature reduction techniques used with

Method Name	Base Classifiers used	Feature reduction technique used	Discussion
CPEMM [31]	NB, RF, SVM, and C4.5 decision tree	PSO	To avoid misclassification and bias, attributes having more than 40% missing data were deleted from the attribute set.
asBagging_FSS [32]	SVM with a Gaussian radial basis function	Clustering with Random projection to generate feature subsets	Asynchronous bagging improves performance with feature selection method. Multiclass dataset is not tested.
RF tree based ensemble [33]	LR, MLP,SVM, RF ,DT and KNN	RF tree	Selected feature size is 27 to 89. Even though many of the datasets are correctly classified 100% of the time, there is still misclassification error, indicating the consequence of overconfidence during correct and incorrect predictions.
eAdaBoost [34]	C4.5, Decision Stump, NB Tree and RF	-	Each sample is weighted based on classification performance. Special decision trees are built to fit high weight samples set. No feature selection is done all dataset is distributed among set of learners.
Ensemble PNN [35]	Probabilistic Neural Network (PNN)	Iterative search margin algorithm Simba and Forward attribute reduction based on neighborhood	PNN is simple and only one parameter tuning is required. Top 25 PNN are selected out of 100 to build ensemble classifier. Only tested for three datasets colon, leukemia and SRBCT. 4 fold CV is used as performance metric.
PTHS[36]	LR, RF, SVM, KNN, NB and DT	Pearson correlation coefficient and Euclidean distance for feature reduction	Parallel optimization is done for parameter tuning of individual learning algorithm and hierarchical selection of models is done with divide and conquer voting system to reduce time complexity. This method does not work well for imbalance datasets
Meta-classifiers [37]	KNN ,SVM and MLP	Filter based nine different ranking algorithms	Evolutionary algorithm is used to select models in ensemble. Farthest first clustering algorithm to generate final prediction. Random partitioning is adopted to test the performance of proposed system.

Table 2 compares various ensemble classifier approaches based on the feature selection method used and the base learner, along with a detailed explanation of the system. To create a homogenous ensemble classifier, ensemble classifier uses the same classifier with varied parameters. The term "heterogeneous ensemble classifier" refers to an ensemble classifier that uses a variety of classifiers. Ensemble PNN [35] is a homogeneous classifier, while [31-34], [36- 37] are heterogeneous classifiers in table 2. Heterogeneous classifiers create diverse ensemble classifiers [38-39].

3. Proposed System

The datasets and methodology used in this paper are discussed in this section.

3.1. Stability Analysis of Ensemble Feature Selection

Model assessment is one of the most important issues in learning theory. Cross-validation is a standard technique in supervised learning for model assessment. It entails training with only a subset of the data and then testing on the remaining data to determine the predictor's estimated risk. In ensemble learning, the stability measure gives an upper bound to cross-validation. The degree of variation in the selected feature subsets given slight changes in the data used to obtain them is known as stability or robustness in feature selection. Ensemble learning has been developed and widely used in recent years to improve the stability of selected feature subsets.

In this part, we'll look at examples of selection stability for various feature selection techniques, as well as how ensemble methods are stable. The effect of ensemble frameworks on feature selection stability has been studied in works such as [40-41]. The majority of these studies, on the other hand, tend to focus on how well these ensembles perform in relation to single selections.

Salman et al. [42] used repeated bootstrap sample generation to explore the behavior of ensemble feature selection in terms of learning accuracy and stability under variation in the aggregation process. The aggregation rules Arithmetic Mean and L2 Norm produce superior stability than any other. In terms of stability performance, Information Gain and MRMR are practically identical.

The methodology used here evaluates the level of stability of the selected features and the prediction performance of the classification models created from them.

Table 3. High dimensional low sample size dataset details used in experiment with dataset name, number of samples as n, number of features as p, number of classes as C_k and subset size used for stability analysis

DN	Dataset Name	n	p	C_k	Subset Size for Stability Analysis
1	COLON	62	2000	2	20
2	Lung	203	12600	2	50
3	Leukemia	72	7129	2	50
4	Prostate	102	12600	2	20
5	MLL	72	12582	3	35
6	Lymphoma	45	4026	3	35
7	SRBCT	83	2308	4	50

The dataset used for experiment is shown in Table 3. Here n is number of instances, p is number of features C_k is number of classes. Here n<<p that the ratio of number of samples with number of features is less than 1. This is known as high dimensional low sample size data. All these seven datasets are high dimensional low sample size imbalanced cancer datasets [43, 44]. These datasets provide a foundation for much of the modern research into cancer and other diseases, guiding experiments, informing clinical decisions, and paving the way for the development of more effective therapies. The subset size selected for stability analysis will be discussed in next section. Simultaneously, the stability of both simple and ensemble ranking is assessed by comparing the 5 subsets similarity analysis.

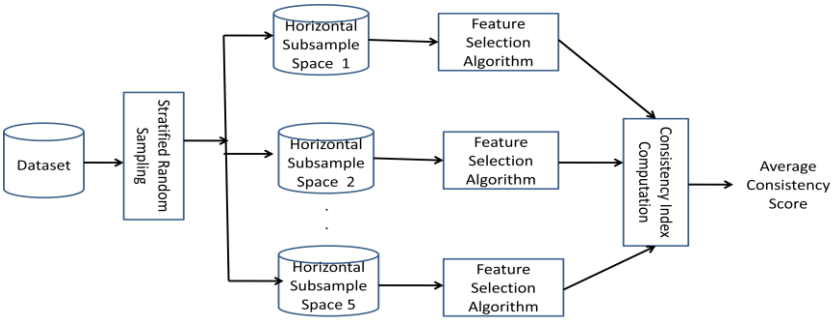


Fig.2. Processing technique adopted for stability analysis of ensemble and simple feature selection

We do this by performing stratified random sampling on the input dataset to generate with 80 percent overlap and 20 percent non overlapped samples in each T distinct training sets. The 5 subsamples are created from stratified random sampling. This sampling technique generates balanced class representation in each subsample space. The detail process is shown in Fig. 2. The two ranking algorithms symmetric uncertainty (SU) and Chi-squares (ChS) are applied for feature selection. Our ensemble feature selection algorithms SU-R and ChS-R [45] are applied to compare stability of feature selected the more similar the subsets are, the more stable the selection method becomes. To assess their degree

of resemblance, we employ an appropriate consistency index [46].

Table 4. Stability measured with Consistency Index in five subsets for SU-R, ChS-R, SU and ChS method on seven high dimensional low sample size datasets

Dataset Number(DN)	Dataset Name	Horizontal partitioned subsets	Stability measured with Consistency index				Improvements	
			ChS-R Stability	SU-R Stability	ChS Stability	SU Stability	Δ_1	Δ_2
1	COLON	SUBSET 1	0.70	0.60	0.39	0.44	0.303	0.152
		SUBSET 2	0.44	0.49	0.24	0.29	0.202	0.202
		SUBSET 3	0.60	0.60	0.34	0.60	0.253	0.000
		SUBSET 4	0.39	0.39	0.24	0.19	0.152	0.202
		SUBSET 5	0.55	0.60	0.29	0.29	0.253	0.303
		AVG	**0.54**	**0.54**	**0.30**	**0.36**	**0.23**	**0.17**
2	Lung	SUBSET 1	0.68	0.64	0.56	0.64	0.119	0.000
		SUBSET 2	0.82	0.88	0.82	0.78	-0.001	0.100
		SUBSET 3	0.92	0.98	0.80	0.80	0.120	0.181
		SUBSET 4	0.84	0.96	0.82	0.84	0.019	0.120
		SUBSET 5	0.80	0.90	0.82	0.68	-0.021	0.221
		AVG	**0.81**	**0.87**	**0.76**	**0.75**	**0.05**	**0.12**
3	Leukemia	SUBSET 1	0.74	0.64	0.50	0.54	0.242	0.101
		SUBSET 2	0.78	0.76	0.60	0.62	0.181	0.141
		SUBSET 3	0.80	0.66	0.60	0.70	0.201	-0.040
		SUBSET 4	0.74	0.76	0.46	0.54	0.282	0.222
		SUBSET 5	0.78	0.72	0.48	0.60	0.302	0.121
		AVG	**0.77**	**0.71**	**0.52**	**0.60**	**0.24**	**0.11**
4	Prostate	SUBSET 1	0.80	0.75	0.70	0.40	0.100	0.351
		SUBSET 2	0.65	0.65	0.65	0.35	0.000	0.300
		SUBSET 3	0.80	0.65	0.60	0.50	0.200	0.150
		SUBSET 4	0.90	0.75	0.60	0.40	0.300	0.351
		SUBSET 5	0.80	0.75	0.65	0.65	0.150	0.100
		AVG	**0.79**	**0.71**	**0.64**	**0.46**	**0.15**	**0.25**
5	MLL	SUBSET 1	0.71	0.38	0.33	0.33	0.382	0.048
		SUBSET 2	0.74	0.58	0.42	0.42	0.327	0.167
		SUBSET 3	0.46	0.33	0.33	0.33	0.124	0.000
		SUBSET 4	0.48	0.54	0.25	0.25	0.236	0.292
		SUBSET 5	0.54	0.29	0.29	0.29	0.251	0.000
		AVG	**0.59**	**0.43**	**0.32**	**0.32**	**0.26**	**0.10**
6	Lymphoma	SUBSET 1	0.74	0.74	0.40	0.35	0.344	0.394
		SUBSET 2	0.80	0.54	0.35	0.40	0.452	0.142
		SUBSET 3	0.74	0.42	0.40	0.30	0.344	0.127
		SUBSET 4	0.80	0.54	0.40	0.50	0.401	0.043
		SUBSET 5	0.86	0.86	0.45	0.50	0.409	0.358
		AVG	**0.79**	**0.62**	**0.40**	**0.41**	**0.39**	**0.21**
7	SRBCT	SUBSET 1	0.73	0.67	0.70	0.65	0.033	0.024
		SUBSET 2	0.71	0.73	0.70	0.70	0.013	0.033
		SUBSET 3	0.67	0.61	0.60	0.60	0.077	0.014
		SUBSET 4	0.61	0.67	0.60	0.65	0.016	0.024
		SUBSET 5	0.69	0.69	0.65	0.60	0.047	0.097
		AVG	**0.68**	**0.67**	**0.65**	**0.64**	**0.04**	**0.04**
Average stability improvement			**71%**	**65%**	**51%**	**50%**	**20%**	**15%**

$$sim_{ij} = \frac{|s_i \cap s_j| - {}^{SJ^2}/_N}{SJ - {}^{SJ^2}/_N} \quad (4)$$

The equation 4 computes consistency index. The size of the subgroups (according to the cutoff threshold) is SJ, and the total number of features is N. The similarity sim_{ij} simply expresses the degree of overlap between subsets. The $|S_i \cap S_j|$ gives intersection of two subsets S_i and S_j . For example there are two sets A={a, b. c ,d, e ,f, g, h } and set B={a, c, e, p, q, r, s} then $|A \cap B| = 3$ i. e. cardinality of common set elements {a, c, e }.

The examples of this score with varying number of features are given in [47]. The resulting similarity scores are then averaged across all pair-wise comparisons to determine the overall degree of similarity among the T subsets and thus the degree of similarity between the T subsets. The selecting process's level of stability is:

$$Stability = sim_{avg} = \sum_{i=1}^{T-1} \sum_{j=i+1}^{T} sim_{ij} \quad (5)$$

This analysis is carried out independently for the simple and ensemble subsets, and different subset sizes are taken into account in both situations. Indeed, the ideal cut-off value may be influenced by the data's properties as well as the ranking criteria used, thus it's worth testing the selection process's resilience across a wide variety of threshold values. Table 3 shows the subset size for each dataset for stability testing, which ranges from 20 to 50. Table 4 illustrates the consistency index estimated with eq. 4 and 5 on five subsets for the SU-R, ChS-R, SU, and ChS methods on seven datasets.

$$\Delta_i = Proposed\ System\ Stability - Existing\ System\ Stability \quad (6)$$

For calculating improvement in stability Δ_1 is computed as ensemble method ChS-R[45] and ChS ranking method for feature selection. Δ_2 is computed as ensemble method SU-R [45] and SU ranking method for feature selection Stability is measured for each horizontal subset produced from the original dataset, and then average stability is calculated. As a result of this experiment, it has been observed that the ensemble technique enhances system stability. According to Table 4, the average ensemble ChS-R methods stability is 71 percent, and the SU-R methods stability is 65 percent with consistency index. These methods improve existing methods stability by an average of 20% and 15%, respectively.

3.2. Proposed Ensemble Classifier

Instead of learning a single classifier, the ensemble principle learns a number of them and then combines their predictions. Individual models have a propensity for being less precise than ensembles. The ensemble approach succeeds because of the diversity in picking individual learners.

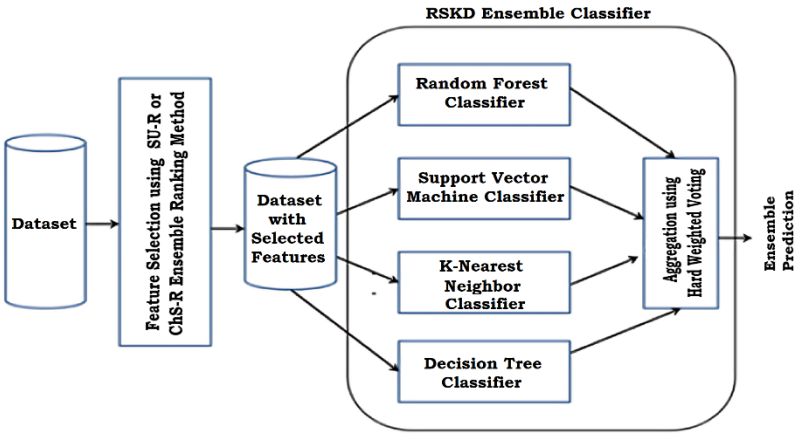


Fig.3. Proposed ensemble classifier RSKD with hard weighted voting

Many models may be trained to teach the learning algorithm how to handle the data. To study diversity of ensemble classification we have implemented RSKD ensemble classifier with RF, SVM, KNN and DT classifiers as a base learner. An ensemble feature selection approach, SU-R and ChS-R, adopted from our earlier work [45], is utilized to select features for training the ensemble classifier. Here we have used high dimensional low sample size datasets mentioned in Table 2.

The detailed working of proposed ensemble classifier RSKD is as shown in figure 3. The SU-R and ChS-R ensemble ranking methods' detailed working is covered in [36] are used for feature selection in first stage. The reduced dataset with selected features is trained using RF, SVM, KNN and DT classifiers. Ensemble classifier is built with these classifiers using weighted hard voting method for result aggregation as per equation1.The classifiers used are RF, SVM, KNN and DT. Random Forest (RF) is an ensemble approach based on decision trees. Support Vector Machine (SVM) generates hyper planes to separate samples belonging to different classes. Here linear kernel is used for generating hyper planes.

K Nearest Neighbor (KNN) works based on proximity of test samples with neighbors instances. Here neighbor count is set to 5 and brute-force search algorithm is used. Decision Tree (DT) represents each node as an attribute test. DT provides data resilience by handling all forms of data well.

The languages Python and R are used in system development and testing. Weka[1] is used for REPtree.

4. Results and Discussions

Performance of the ensemble classifier RSKD on seven high-dimensional datasets is shown in Table 5. The accuracy is used as performance metric and calculated as per equation 7. It is calculated as addition of True positive (TP) and True Negative (TN) divided by total number of instances. The accuracy is converted to percentage accuracy by multiplying it with 100.

$$\%Accuracy = \frac{TN + TP}{Total\ number\ of\ instances} * 100 \tag{7}$$

Table 5 Accuracy of Ensemble Classifier RSKD with 5 fold cross validation for ChS-R and SU-R ensemble feature selection methods on seven high dimensional low sample size cancer datasets

DN	Dataset Name	Feature Selection Methods	Proposed Ensemble Classifier RSKD	Existing system accuracy			Improvement over existing methods		
				Adaboost	GradientBoost	REPTree[1]	Δ_3	Δ_4	Δ_5
1	COLON	SU-R	92	70.2	85	75.8	21.8	7	16.2
		ChS-R	92.11	77.2	76.92	80.64	14.91	15.19	11.47
2	Lung	SU-R	100	97.4	96.87	97.87	2.6	3.13	2.13
		ChS-R	100	98.1	98.1	96.79	1.9	1.9	3.21
3	Leukemia	SU-R	99.68	97.1	96.5	98.01	2.58	3.18	1.67
		ChS-R	99.63	89.4	92.4	90.27	10.23	7.23	9.36
4	Prostate	SU-R	94.44	82.3	82.4	81.37	12.14	12.04	13.07
		ChS-R	94.73	82.7	84.4	79.41	12.03	10.33	15.32
5	MLL	SU-R	95.74	61.7	93.33	83.33	34.04	2.41	12.41
		ChS-R	95.45	88	87.8	90.27	7.45	7.65	5.18
6	Lymphoma	SU-R	100	98.6	98.8	87.09	1.4	1.2	12.91
		ChS-R	95.33	94.8	90.66	95.16	0.53	4.67	0.17
7	SRBCT	SU-R	98.5	75.2	87.1	87.95	23.3	11.4	10.55
		ChS-R	89.85	61.8	69.5	72.28	28.05	20.35	17.57
Average			**96.25**	**83.89**	**88.56**	**86.87**	**12.35**	**7.69**	**9.37**
Average SU-R			97.19	83.21	91.43	87.35	13.98	5.77	9.85
Average ChS-R			95.30	84.57	85.68	86.40	10.73	9.62	8.90

[1]https://weka.sourceforge.io/doc.dev/weka/classifiers/trees/REPTree.htm

The proposed ensemble classifier has an average performance accuracy of 96.25 percent. The average performance accuracy of the SU-R feature selection approach ensemble classifier is 97.19 percent, whereas the average performance accuracy of the ChS-R ensemble classifier is 95.30 percent. The system is compared to existing state-of-the-art methodologies Adaboost, Gradient Boost, and [1]REPTree algorithms by computing Δ_3 , Δ_4 and Δ_5 are computed as per equation 6. Here Δ_3 shows improvement over Adaboost , Δ_4 shows improvement over Gradient Boost and Δ_5 improvement over REPTree algorithms. The proposed ensemble classifier RSKD exhibits an average improvement of 12.35 percent, 7.69 percent, and 9.37 percent calculated as Δ_3 , Δ_4 and Δ_5, respectively. The SRBCT dataset shows the most improvement, followed by COLON. The lung dataset shows the least average improvement, followed by lymphoma. It has been discovered that the accuracy of ensemble classifiers is influenced by feature reduction strategies employed before them. The average ensemble classifier accuracy with ChS-R as feature selection is 95.30 while when SU-R feature selection method is used it is 97.19. The RSKD ensemble classifier achieves better accuracy than existing state-of-the-art methods. This is due to the use of SU-R and ChS-R ensemble feature selection techniques for stable feature selection and the diverse combination of classifiers, which enhances its generalization ability. These factors collectively contribute to improved system performance.

We have compared our system with existing ensemble classifiers like as Bagging_FSS[32], SRKNN[48], MF-GE[49]and eAdaBoost with DSC[34]. The comparison ensemble classifiers are shown in Table 6. The table has no entries between the proposed ensemble classifier and the current for some cells since the same datasets were not used in the comparison methods. The graph has been plotted as indicated in figure 4 in order to observe the improvement. It has been found that the current ensemble classifier outperforms all other state-of-the-art techniques.

Table 6. Comparison of proposed RSKD Ensemble Classifier accuracy with existing state of art ensemble classifiers asBagging_FSS, SRKNN, MF-GE and eAdaBoost with DSC

Dataset	Proposed RSKD with SU-R	Proposed RSKD with ChS-R	asBagging_FSS [40]	SRKNN [48]	MF-GE [49]	eAdaBoost with DSC[42]
COLON	92	**92.11**	84.68	80	76.98	-
Lung	**100**	**100**	72.82	93	-	84.38
Leukemia	**99.68**	99.63	-	89	96.23	-
Prostate	94.44	**94.73**	-	91	-	91.18
MLL	**95.74**	95.45	-	-	91.08	-
Lymphoma	**100**	95.33	-	98	-	-
SRBCT	**98.5**	89.85	-	-	-	

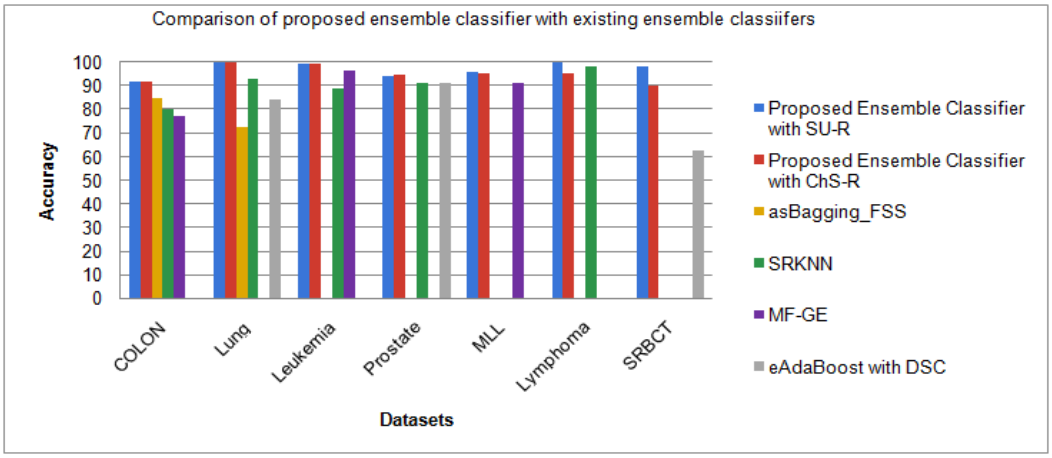


Fig.4. Comparison of proposed RSKD ensemble classifier with existing ensemble classifiers

Significance Test with Wilcoxon Signed-ranks Test

A pairwise Wilcoxon Signed-Ranks test [50], a nonparametric statistical test, is employed to assess the significance of the proposed method compared to other methods. The test is conducted using Excel, with all experiments performed at a significance level of 0.05. The results indicate that the proposed RSKD with SU-R and RSKD ChS-R classifiers show significant differences when compared to the Adaboost, GradientBoost, REPTree. The significance test conducted with hybrid methods and other state-of-the-art ensemble methods, including Bagging_FSS [40], SRKNN [48], MF-GE [49], and eAdaBoost with DSC [42], also reveals statistically significant differences. The results of the pairwise Wilcoxon Signed-Ranks test confirm that the proposed ensemble classifiers methods outperform existing state-of-the-art methods with statistically significant differences at a significance level of 0.05. These findings demonstrate the robustness and effectiveness of the proposed approaches in comparison to established techniques.

5. Conclusions

This work presented a stable ensemble classifier RSKD implemented for high dimensional low sample cancer datasets. When using ensemble feature selection, a smaller feature set with higher accuracy is produced. Any feature selection method's stability is a study topic. When compared to single measure feature selection, the ensemble feature selection methods ChS-R and SU-R demonstrate good stability. This stability increase is significant, increasing stability by 15 to 20 percent. There is no single classifier that performs consistently for high dimensional low sample size datasets. In this study, existing ensemble classifiers were compared with the RSKD ensemble classifier, which provides greater accuracy improvement for this purpose. The proposed RSKD classifier can help to identify critical genes or molecular markers associated with cancer types, providing insights for further biological validation. The datasets employed in this study represent diverse cancer types with varying molecular profiles, demonstrating the generalizability of the proposed ensemble classifier. Its ability to handle high-dimensional data, combined with the feature selection strengths of SU-R and ChS-R, ensures robust and interpretable results across different cancer types. By facilitating biomarker discovery, improving diagnostic accuracy, and supporting treatment personalization, the classifier holds great promise for advancing both cancer research and clinical oncology. Dynamic ensemble selection may be used in the future for research.

References

[1] Borja Seijo-Pardo, Verónica Bolón-Canedo, and Amparo Alonso-Betanzos,"Testing Different Ensemble Configurations for Feature Selection", Neural Process. Lett. 46, 3 (December 2017), pp. 857–880, 2017. https://doi.org/10.1007/s11063-017-9619-1

[2] Barbara Pes, Nicoletta Dessì, Marta Angioni, Exploiting the ensemble paradigm for stable feature selection: A case study on high-dimensional genomic data, Information Fusion, Volume 35, 2017, Pages 132-147, ISSN 1566-2535, https://doi.org/10.1016/j.inffus.2016.10.001.

[3] V. Nikulin, On the homogeneous ensembling via balanced subsets combined with wilcoxon-based feature selection., in: In: Yao J. et al. (eds) Rough Sets and Current Trends in Computing. RSCTC 2012. Lecture Notes in Computer Science, vol 7413., Springer, Berlin, Heidelberg, 2012.

[4] V. Bol´on-Canedo, N. S´anchez-Maro˜no, A. Alonso-Betanzos, Data classification using an ensemble of filters, Neurocomputing 135 (2014) 13–20.

[5] V. Bol´on-Canedo, N. S´anchez-Maro˜no, A. Alonso-Betanzos, An ensemble of filters and classifiers for microarray data classification, Pattern recognition 45 (2012) 531–539.

[6] B. Seijo-Pardo, I. Porto-D´ıaz, V. Bol´on-Canedo, A. Alonso-Betanzos, Ensemble feature selection: Homogeneous and heterogeneous approaches, Knowledge-Based Systems 114 (2017) 124–139.

[7] B. Seijo-Pardo, V. Bol´on-Canedo, A. Alonso-Betanzos, On developing an automatic threshold applied to feature selection ensembles, Information Fusion (2019) 1227–245.

[8] M. Haque, N. Noman, R. Berretta, P. Moscato, Heterogeneous ensemble combination search using genetic algorithm for class imbalanced data classification, PLoS One 11 (2016) e0146116.

[9] D. Guru, M. Suhil, S. Pavithra, G. Priya, Ensemble of feature selection methods for text classification: An analytical study., in: In: Abraham A., Muhuri P., Muda A., Gandhi N. (eds) Intelligent Systems Design and Applications. ISDA 2017. Advances in Intelligent Systems and Computing, vol 736., Springer, Cham,2018.

[10] A. Ben Brahim, M. Limam, Ensemble feature selection for high dimensional data: a new method and a comparative study, Advances in Data Analysis and Classification 1–16doi:10.1007/s11634-017-0285-y.

[11] B. Seijo-Pardo, I. Porto-Díaz, V. Bolón-Canedo, A. Alonso-Betanzos, Ensemble feature selection: Homogeneous and heterogeneous approaches, Knowledge-Based Systems, Volume 118, 2017, pp. 124-139, ISSN 0950-7051, https://doi.org/10.1016/j.knosys.2016.11.017.

[12] F. Moreno-Seco, J. M. Iñesta, P. J. de León, and L. Micó, "Comparison of classifier fusion methods for classification in pattern recognition tasks," *Lecture Notes in Computer Science*, pp. 705–713, 2006.

[13] Clark, P. and Boswell, R., "Rule induction with CN2: Some recent improvements." In Proceedings of the European Working Session on Learning, pp. 151-163, Pitman, 1991.

[14] Brodley, C. E., Automatic selection of split criterion during tree growing based on node selection. In Proceedings of the Twelth International Conference on Machine Learning, 73-80 Taho City, Ca. Morgan Kaufmann, 1995.

[15] Schapire, R.E. The strength of weak learnability. *Mach Learn* 5, 197–227 (1990). https://doi.org/10.1007/BF00116037

[16] Wolpert, David H. and William G. Macready. "Combining Stacking With Bagging To Improve A Learning Algorithm." (1996).

[17] A. K. Seewald and J. Fürnkranz, "An Evaluation of Grading Classifiers," Advances in Intelligent Data Analysis, pp. 115–124, 2001, doi: 10.1007/3-540-44816-0_12.

[18] Chan P. K. and Stolfo, S. J., Toward parallel and distributed learning by metalearning, In AAAI Workshop in Knowledge Discovery in Databases, pp. 227-240, 1993.

[19] Chan P.K. and Stolfo, S.J., A Comparative Evaluation of Voting and Metalearning on Partitioned Data, Proc. 12th Intl. Conf. On Machine Learning ICML-95, 1995.

[20] Chan P.K. and Stolfo S.J, On the Accuracy of Meta-learning for Scalable Data Mining, J. Intelligent Information Systems, 8:5-28, 1997.

[21] Vipin Kumar and Sonajharia Minz. 2015. Multi-view Ensemble Learning: A Supervised Feature Set Partitioning for High Dimensional Data Classification. In *Proceedings of the Third International Symposium on Women in Computing and Informatics* (*WCI '15*). Association for Computing Machinery, New York, NY, USA, 31–37. https://doi.org/10.1145/2791405.2791443

[22] Kumar, V., Minz, S. Multi-view ensemble learning: an optimal feature set partitioning for high-dimensional data classification. *Knowl Inf Syst* 49, 1–59 (2016). https://doi.org/10.1007/s10115-015-0875-y

[23] Lopez-Garcia, P., Masegosa, A.D., Osaba, E. *et al.* Ensemble classification for imbalanced data based on feature space partitioning and hybrid metaheuristics. *Appl Intell* 49, 2807–2822 (2019). https://doi.org/10.1007/s10489-019-01423-6

[24] Yongjun Piao, Minghao Piao, Cheng Hao Jin, Ho Sun Shon, Ji-Moon Chung, Buhyun Hwang, Keun Ho Ryu, "A New Ensemble Method with Feature Space Partitioning for High-Dimensional Data Classification", *Mathematical Problems in Engineering*, vol. 2015, Article ID 590678, 12 pages, 2015. https://doi.org/10.1155/2015/590678

[25] D.B. Skillicorn, S.M. McConnell, Distributed prediction from vertically partitioned data, Journal of Parallel and Distributed Computing, Volume 68, Issue 1, 2008, Pages 16-36, ISSN 0743-7315, https://doi.org/10.1016/j.jpdc.2007.07.009.

[26] Mohammed, A.M., Onieva, E., Woźniak, M. (2020). Vertical and Horizontal Data Partitioning for Classifier Ensemble Learning. In: Burduk, R., Kurzynski, M., Wozniak, M. (eds) Progress in Computer Recognition Systems. CORES 2019. Advances in Intelligent Systems and Computing, vol 977. Springer, Cham. https://doi.org/10.1007/978-3-030-19738-4_10

[27] Kusiak, A., Decomposition in Data Mining: An Industrial Case Study, IEEE Transactions on Electronics Packaging Manufacturing, Vol. 23, No. 4, pp. 345-353, 2000.

[28] Vega-Pons, S., Ruiz-Shulcloper, J. (2009). Clustering Ensemble Method for Heterogeneous Partitions. In: Bayro-Corrochano, E., Eklundh, JO. (eds) Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. CIARP 2009. Lecture Notes in Computer Science, vol 5856. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-10268-4_56

[29] Arkadiy Dushatskiy, Tanja Alderliesten, and Peter A. N. Bosman. 2021. A novel surrogate-assisted evolutionary algorithm applied to partition-based ensemble learning. In Proceedings of the Genetic and Evolutionary Computation Conference

(GECCO '21). Association for Computing Machinery, New York, NY, USA, 583–591. https://doi.org/10.1145/3449639.3459306
[30] Xiaoli Zhang Fern and Carla E. Brodley. 2004. Solving cluster ensemble problems by bipartite graph partitioning. In Proceedings of the twenty-first international conference on Machine learning (ICML '04). Association for Computing Machinery, New York, NY, USA, 36. https://doi.org/10.1145/1015330.1015414
[31] T. R. Sivapriya, A. R. Nadira Banu Kamal, P. Ranjit Jeba Thangaiah, "Ensemble Merit Merge Feature Selection for Enhanced Multinomial Classification in Alzheimer's Dementia", *Computational and Mathematical Methods in Medicine*, vol. 2015, Article ID 676129, 11 pages, 2015. https://doi.org/10.1155/2015/676129
[32] H. Yu and J. Ni, "An Improved Ensemble Learning Method for Classifying High-Dimensional and Imbalanced Biomedicine Data," in *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 11, no. 4, pp. 657-666, July-Aug. 2014, doi: 10.1109/TCBB.2014.2306838.
[33] Dagnew, G. and Shekar, B.," Ensemble learning-based classification of microarray cancer data on tree-based features", Cogn. Comput. Syst, 3: 48-60. https://doi.org/10.1049/ccs2.12003 (2021)
[34] S. Dinakaran and P. Ranjit Jeba Thangaiah," Ensemble Method of Effective AdaBoost Algorithm for Decision Tree Classifiers", International Journal on Artificial Intelligence Tools 26(03), https://doi.org/10.1142/S0218213017500075(2017)
[35] Shu-Lin Wang, Xueling Li, Shanwen Zhang, Jie Gui, De-Shuang Huang, "Tumor classification by combining PNN classifier ensemble with neighborhood rough set based gene reduction", Computers in Biology and Medicine 40(2),pp. 179-189, ISSN 0010-4825, https://doi.org/10.1016/j.compbiomed.2009.11.014 (2010).
[36] Wei, Leyi et al. "A novel hierarchical selective ensemble classifier with bioinformatics application." *Artificial intelligence in medicine* vol. 83 (2017): 82-90. doi:10.1016/j.artmed.2017.02.005
[37] Kim, KJ., Cho, SB. Meta-classifiers for high-dimensional, small sample classification for gene expression analysis. *Pattern Anal Applic* 18, 553–569 (2015). https://doi.org/10.1007/s10044-014-0369-7
[38] M. Paz Sesmero, José Antonio Iglesias, Elena Magán, Agapito Ledezma, Araceli Sanchis, Impact of the learners diversity and combination method on the generation of heterogeneous classifier ensembles, Applied Soft Computing, Volume 111, 2021, 107689, ISSN 1568-4946, https://doi.org/10.1016/j.asoc.2021.107689.
[39] Tien Thanh Nguyen, Mai Phuong Nguyen, Xuan Cuong Pham, Alan Wee-Chung Liew, Heterogeneous classifier ensemble with fuzzy rule-based meta learner, Information Sciences, Volume 422, 2018, pp. 144-160, ISSN 0020-0255, https://doi.org/10.1016/j.ins.2017.09.009.
[40] Pes B. Ensemble feature selection for high-dimensional data: a stability analysis across multiple domains. Neural Comput Appl. 2020;32(10):5951–73, https://doi.org/10.1007/s00521-019-04082-3.
[41] Xiang, Feng, Yulong Zhao, Meng Zhang, Ying Zuo, Xiaofu Zou, and Fei Tao. "Ensemble learning-based stability improvement method for feature selection towards performance prediction." Journal of Manufacturing Systems 74 (2024): 55-67. https://doi.org/10.1016/j.jmsy.2024.03.001
[42] Salman, R., Alzaatreh, A. & Sulieman, H. The stability of different aggregation techniques in ensemble feature selection. *J Big Data* 9, 51 (2022). https://doi.org/10.1186/s40537-022-00607-1
[43] https://archive.ics.uci.edu/ml/datasets.php
[44] https://csse.szu.edu.cn/staff/zhuzx/Datasets.html
[45] Sumant, A.S., Patil, D. Ensemble Feature Subset Selection: Integration of Symmetric Uncertainty and Chi-Square techniques with RReliefF. J. Inst. Eng. India Ser. B 103, 831–844 (2022). https://doi.org/10.1007/s40031-021-00684-5
[46] Kuncheva LI (2007) A stability index for feature selection. In: 25th IASTED international multi-conference: artificial intelligence and applications, ACTA Press, pp 390–395.
[47] Cannas LM, Dessı` N, Pes B (2013) Assessing similarity of feature selection techniques in high-dimensional domains. Pattern Recognit Lett 34(12):1446–1453
[48] M. Paz Sesmero, José Antonio Iglesias, Elena Magán, Agapito Ledezma, Araceli Sanchis, Impact of the learners diversity and combination method on the generation of heterogeneous classifier ensembles, Applied Soft Computing, Volume 111, 2021, 107689, ISSN 1568-4946, https://doi.org/10.1016/j.asoc.2021.107689.
[49] Tien Thanh Nguyen, Mai Phuong Nguyen, Xuan Cuong Pham, Alan Wee-Chung Liew, Heterogeneous classifier ensemble with fuzzy rule-based meta learner, Information Sciences, Volume 422, 2018, pp. 144-160, ISSN 0020-0255, https://doi.org/10.1016/j.ins.2017.09.009.
[50] Arora, P., Mishra, A., & Malhi, A. (2024). An Ensemble Machine Learning Method Highlights Possible Parkinson's Disease Genes and Accessing Performance of Re-sampling Techniques. SN Computer Science, 5(5), 483

Authors' Profiles

Archana S. Vaidya received a Ph.D. degree in computer engineering from Savitribai Phule Pune University (SPPU), Pune, Maharashtra, India in 2023. She received her master's degree in computer engineering from V.J.T.I., Mumbai University (INDIA) in 2010 and bachelor's degree in computer engineering from Walchand College of Engineering Sangli Shivaji University (INDIA) in 2002. Her areas of interest are parallel computing and machine learning and data analytics. She has teaching experience of 22 years. She is a life member of ISTE and IAENG. She has published 9 Scopus-indexed papers and 40+ journal papers. She can be contacted at archana.vaidya@ges-coengg.org.

Dipak V. Patil received a B.E. degree in computer engineering in 1998 from the University of North Maharashtra India and an M.Tech. degree in computer engineering in 2004 from Dr. Babasaheb Ambedkar Technological University, Lonere, India. He has done a Ph.D. degree from S.R.T.M. University, Nanded. Currently, he is a professor and head of the Computer Engineering Department at GES R.H. Sapat College of Engineering, Management Studies and Research, Nashik, India which is affiliated with the Savitribai Phule Pune University, Pune, and Maharashtra, India 422002. He can be contacted at email: dipakvpatil17@gmail.com.

How to cite this paper: Archana Suhas Vaidya, Dipak V. Patil, "RSKD Ensemble Classifier with Stable Ensemble Feature Selection for High Dimensional Low Sample Size Cancer Datasets", International Journal of Information Technology and Computer Science(IJITCS), Vol.17, No.2, pp.49-59, 2025. DOI:10.5815/ijitcs.2025.02.05