International Journal of Applied Mathematics

Volume 38 No. 5, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

ENSEMBLE CLUSTERING OF FEATURE RANKS FOR FEATURE SELECTION

Swetha T[1*], Sakthivel G[1†] and Sandhya Rani D[2†]

[1*]Department of Electronics and Instrumentation Engineering, Annamalai University, Annamalainagar, Chidambaram, 608002, Tamil Nadu, India.

[2]Department of Computer Science and Engineering, CVR College of Engineering, Ibrahimpatnam, Hyderabad, 501510, Telangana, India.

*Corresponding author(s). E-mail(s): tungaturthiswetha1234@gmail.com; Contributing authors: gsauei@gmail.com; d.sandhyarani@cvr.ac.in;

[†]These authors contributed equally to this work.

Abstract

Selecting a subset of features poses an NP hard problem, necessitating the development of computationally efficient algorithms to identify nearly optimal feature subsets that enhance classifier performance. High-dimensional data, featured with a numerous features, and extensive datasets present significant challenges to feature subset selection. Key concerns include the extensibility of the feature selection methods in terms of high-dimensional data accuracy and processing time for big datasets.

To address these problems, we devised a feature selection method based on ensemble clustering. The literature offers numerous computationally efficient greedy feature ranking techniques, each ranking features differently. A Ensemble among these varied rankings can yield a feature ranking method whose time will be less and performance will be high. This research aims to develop effective algorithms that can deal with both high dimensional and also small datasets.

Our contributions include the design of Feature Ranking based on Ensemble Clustering (FREC), a rapid and expandable method for feature selection that leverages existing feature ranking algorithms. Implementation results demonstrate that FREC noticeably outperforms several current methods found in the literature on both small and high dimensional datasets.

Keywords: Feature subset selection, Ensemble clustering, Feature ranking method.

1 Introduction

Feature selection is a critical step in data analysis, aiming to identify a subset of characteristics strongly correlated with the class variable. This process typically involves navigating a vast search space encompassing all possible combinations of data features. The increasing dimensionality of datasets in real-world applications, such as gene expression databases, has

Received: August 08 2025

International Journal of Applied Mathematics
Volume 38 No. 5, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

significantly amplified the complexity and time required to pinpoint the most effective feature subset from high-dimensional data.

Feature ranking offers a straightforward approach to feature subset selection. Most feature ranking algorithms adopt a greedy strategy, which makes them more computationally efficient compared to other methods. Numerous feature ranking algorithms are readily available, each producing a ranked list of features from most to least important. These algorithms employ diverse ranking criteria, such as feature weight, information-theoretic measures, and statistical measures, leading to varied feature orderings across different algorithms.

This paper introduces a novel feature selection algorithm called Feature Ranking based on Ensemble Clustering (FREC). FREC leverages the concept of 'Ensemble clustering' by integrating insights from multiple feature ranking algorithms. For large-scale applications, a parallel version of FREC, known as Hybrid Feature Selection (HFS), is also proposed. As highlighted by Barbara Pes, HFS incorporates diversity at both the data and algorithm levels. Data-level diversity is achieved by partitioning the dataset, while algorithm-level diversity is introduced by applying non similar feature rankers to every partition.

1.1 Motivation of the proposed method

Many features in the large dimensional data sets could be redundant and unim- portant. The performance of the classifier may also be impacted by such characteristics. To find the most pertinent features, a variety of feature ranking algorithms are available. In the process of extracting the most pertinent features and eliminating redundant ones, Ensemble clustering is used.

The remainder of the paper is organised as follows. Section 2 deals with some approximation techniques utilised in the literature and ensemble clustering also introduced here. Section 3 presents related work of these techniques. The suggested algorithm, FREC, is introduced in section 4. Complexity analysis and experimental results are described in section 5. Section 6 summarizes the conclusions.

2 Ensemble clustering background

In this section we have introduced ensemble clustering and various methods

available in literature based on this method.

2.1 Ensemble Clustering

Data objects are grouped into groups using clustering so that those in the same cluster are very similar to one another and those in different clusters are not. [2]. Nevertheless, there are a number of clustering algorithms in the literature that have certain disadvantages, including the use of similarity mea- sures, sensitivity to initial parameter settings in K-means, the lack of privious knowledge regarding the number of total clusters, increase in the time complexity with the increase of dimensions increase, etc. [3]. Furthermore, the data may be clustered differently by different clustering algorithms. Clustering output is therefore unstable. Cluster ensembling, also known as clustering aggregation or

Received: August 08 2025

International Journal of Applied Mathematics
Volume 38 No. 5, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Ensemble clustering [7], can be utilised to solve this issue. Finding the optimal partitioning that matches with several input clusterings or partitionings is the primary objective of Ensemble clustering. The input clusterings or partitionings can be produced by either executing alternative clustering algorithms [**10**] or repeatedly running the same clustering algorithm with varied initial settings of input parameters (different K-value in K-means). It is NP-hard to find Ensemble among potentially exponentially many input clusterings. As a result, numerous heuristics have been developed in the literature [**12**] to solve this issue.

Out of several clustering solutions, Ensemble clustering finds one that is superior to the others [4, 5] clusterings. It incorporates results of clustering from same data from numerous sources, according to [6, 7]. This has grown in popularity because of its dependability and efficiency. Conditional clustering, outlier analysis, embedding of graphs, and other issues have been resolved with the help of the Ensemble clustering framework. Liu et al., [8], have employed a Ensemble approach for unsupervised feature selection in recent years. The authors have employed a Ensemble framework for spectral ensemble clustering in the publication [9]. Because Ensemble clustering uses cluster labels produced by various clustering or partitioning algorithms as input rather than the original base data, it can be thought of as privacy-preservation clustering. We solve the feature subset selection problem using this framework.

2.2 Ensemble clustering algorithms

A number of algorithms, such the *Average linkage* algorithm and the *Fur- thest* algorithm, are available in the literature to reach a Ensemble. With an $O(n^2(logn+k))$ time complexity, the former is an agglomerative method, where 'K' represents the number of input partitionings and 'n' represents the number of instances. Numerous approximation techniques, such as simulated annealing, BestOneElementMove (BOEM), MajorityRule(MR), CC-Pivot, BestOfK(BOK), and CCLP-Pivot algorithm, are accessible in the literature [**12**]. Each of these algorithms uses a dissimilarity matrix to determine the Ensemble clustering. The temporal complexity of all these approximation methods, with the exception of BOK, is at least $O(n^2)$. BOK has a linear time complexity of $O(K^2n)$ with 'K' input clusterings and 'N' instances.

2.2.1 BestOfK algorithm

Various metrics are available in the literature to assess how distinct or similar two partitionings or clusterings C_1 and C_2 are. Normalised Mutual Informa- tion (NMI), an entropy-based similarity metric, was proposed by Strhel and Ghosh [6]. Other widely used metrics from the literature include the Adjusted Rand Index (ARI), Quality Partition Index (QPI)

$$d(C , C) = (q_1 + r) \; or \; {}^{n} - (p_2 + s) \qquad (1)$$

Given the two clusterings C_1 and C_2, 'p' denotes the number of object pairs that overlapped in both clusterings, and 'q' denotes the number of object pairs that overlapped in C_1 but

Received: August 08 2025

International Journal of Applied Mathematics
Volume 38 No. 5, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

not in C_2.[40] The number of pairs of objects that co- occurred in C_2 but not in C_1 is represented by 'r', while the number of pairs of objects that did not co-occur in both clusterings is represented by 's'.

Ensemble clustering states that the best clustering is the one with the low- est SDD score among all the other clusterings [**13**]. In other words, given input clusterings (C_1, C_2, C_3,….C_k), we must discover a clustering or partitioning

C^* such that

$$C^* = \arg\min_{C} \sum_{i=1}^{k} d(C_i, C) \qquad (2)$$

2.3 Feature ranking algorithms

The literature's conventional filter-based feature ranking techniques χ^2 chi- squared forIn addition to feature weights, feature rankings are generated using [10], Information Gain(IG) [11], Our algorithms have been coded in Python, and the Weka tool has been utilised to generate clusters and improve classification accuracy. A few classifiers have been run using R-studio. The Windows OS platform has been used for the entire project.

3 Related Work

Recent literature has addressed a wide range of feature selection algorithms [**11**

][**12**]. This section discusses a few key feature subset selection techniques.

3.1 Feature selection based on genetic algorithms

One of the widely used feature selection techniques that has been effectively applied is the genetic algorithm

3.2 Feature selection based on redundancy and relevance

Relevant elements can be found using a few conventional literary techniques. One best-known example, presented by Kira and Rendell is Relief [12], is based on the weight of the feature.

Next, Yu and Liu [13] categorise the feature sets as 'weakly relevant' but ' non-redundant' features, 'redundant', 'irrelevant', and 'strongly relevant' features. According to them, features that are non-redundant but highly significant must be included in an ideal feature subset.

Methods that deal with redundant and irrelevant features include NMIFS [14], mRMR [15], FCBF [16], SBC [17], and FAST [18]. Symmetric uncertainty (SU) has been employed by Lei et al. [19] as a metric to identify pertinent features.

3.3 Feature subset selection methods based on graphs

Since this method shows the relationship between feature vectors or features, one best way to determine the ideal subset of features is to use a graph to represent the feature space and

Received: August 08 2025

International Journal of Applied Mathematics
Volume 38 No. 5, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

identify the feature groups. The three-step FAST method was proposed by Song et al., [18]. The Symmetric Uncertainty measure is used in the first phase to eliminate features that aren't important. The second stage then uses the feature graph to create a minimal spanning tree (MST). In order to choose the representative features, the minimum spanning tree is finally divided. The algorithm's time complexity is $O(Mlog^2M)$, where M is the number of features. Graph-based feature selection techniques like Laplacian scoreZhang et al [20] offer an information theoretic strategy for feature selection based on hypergraphs. Bandopadhyay et al,. [22] Mandal et al. [21]. proposed a graph based approach that is an unsupervised feature selection method. This method removes redundant features from the whole set of features.

3.4 Feature ranking methods with Ensemble approach

Some well-known techniques based on the ensemble approach are found in the literature, such as FRMV [23], EFR [24], and others. FRMV is a multi- perspective unsupervised feature ranking system. The author employed the symmetric uncertainty (SU) measure to rank the features, combining several feature rankers from different subsets of the same dataset into a single con- sensus. The EFR approach was proposed by Jong et al. for feature rating. By merging feature rankings from separate outputs of ROC-based genetic learner, the authors of this article employ the ensemble technique for feature ranking.

3.5 Research gaps

To identify the feature subset that is optimal, the majority of conventional feature selection methods require a considerable amount of learning time. Even if classifier accuracy is excellent, the main issues with most feature selection techniques, such as genetic algorithms, are scalability and feature reduction. Recent research has introduced parallelizable methods, such as fuzzy rough set techniques, with the goal of minimising run time and optimising memory utilisation. The number of characteristics is not significantly reduced by several of these techniques. Finding the ideal feature subset while maintaining high classifier performance becomes extremely difficult, particularly when working with big datasets that contain a very high number of redundant and irrelevant features.

4 Feature rankers based feature subset selection framework

For each ranking algorithm *RA*, the weight of the feature is used as the sole dimension while doing K-means clustering. The cluster labels are obtained using K-means clustering. When M denotes the number of features and *n* denotes the number of ranking algorithms, an input file with rows of *M* and columns of *n* is produced. The cluster labels that the K-means algorithm allocated to the features are listed in each column. The Ensemble clustering algorithm receives this matrix as input. The BestOfK Ensemble algorithm is used to determine the optimal clustering. Last but not least, features from the top cluster are selected since their combined weights are the highest. The FREC method is depicted in Figure 1.

Received: August 08 2025

International Journal of Applied Mathematics

Volume 38 No. 5, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

The detail description of FREC algorithm is given in Algorithm 1.

O(nNM) is the amount of time needed to calculate using n different feature ranks with M features. If I denots the iteration number, the time needed to split M number of features using K-means clustering approaches is $O(KIM)$. The optimal partitioning will be found in O(n^2M) time using BestOfK (BoK) Ensemble clustering. The total time complexity of the method is therefore O(($nNM + nKMI + n^2M$)) $= O(N + KI + n)nM = O(NM)$. With a time complexity of O(NM^2), FREC is faster and more scalable than our prior approach, FSSCC[25]. Particularly when microarray datasets are used where M is very high compared to N, this time is extremely high.

4.0.1 Variations of the FREC

Since each dataset has a varied amount of features in them, we divide the data sets into three groups: low feature group, medium feature group and high feature group data sets. We then apply FREC methods to each set, slightly altering them.

1. **Datasets with low dimensions (less than 100 features):** For low dimension data sets like Wine-8 to Sonar-60, the FREC algorithm is

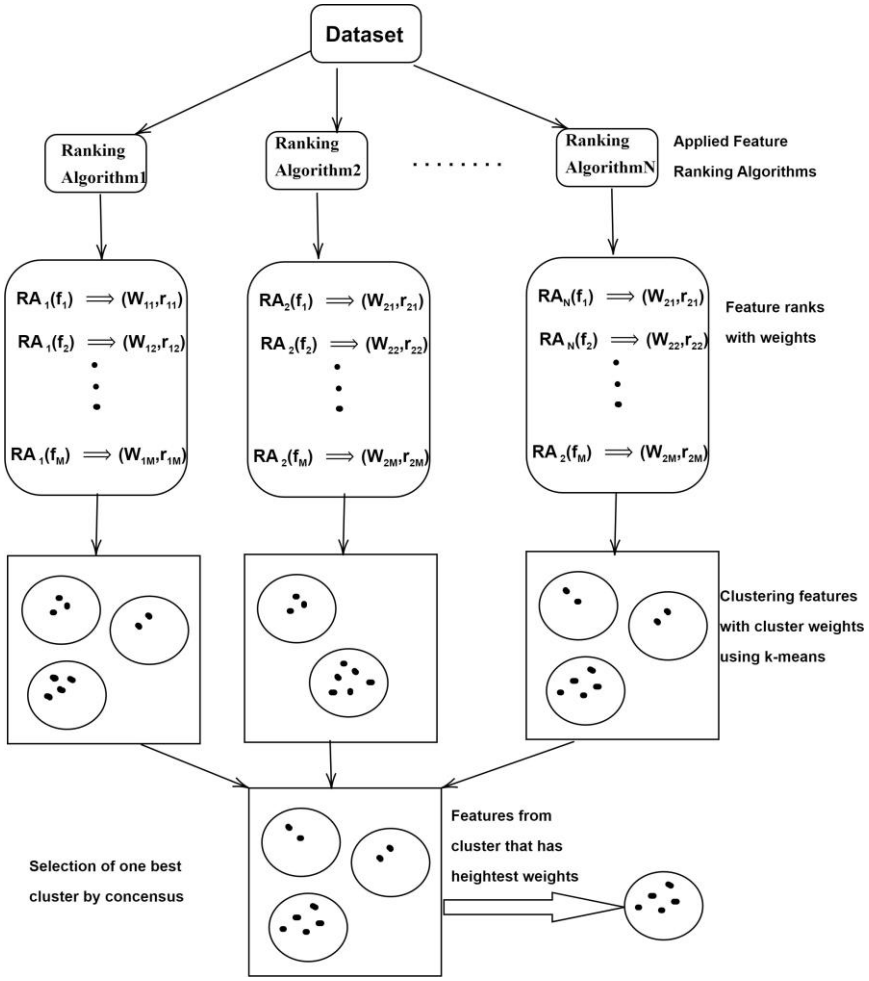

Fig. 1: Ensemble clustering using feature rankers

Received: August 08 2025

International Journal of Applied Mathematics
Volume 38 No. 5, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

utilised. The best cluster is the one which has maximum sum of weights will be chosen as the final feature subset.

2. Medium sized datasets with 100–1000 features: The FREC approach is used initially for medium set of features or dimensions datasets which are having more than 100 features like Musk1 and Musk2 (168 features). Then, the K-means algorithm is applied repetitively on the first iteration output cluster to limit the subset of features. Following each iteration, the classifier's accuracy is assessed. This process is repeated until using the chosen characteristics causes the classifier's accuracy to stop declining.

Algorithm : Feature ranking- based on ensemble clustering (FREC)

Input:

- Dataset X = (x, t), where N is the number of instances.

- Feature subset F = {f_1, f_2, ..., f_m}.

- Target vector t.

- Feature weight set W = {w_1, w_2, ..., w_m}.

- A collection of Feature Ranking Algorithms (FRA): { Gain Ratio, Information Gain, OneR, ReliefF, Symmetric Uncertainty}.

Output:

- Final Feature Subset (FS)

Procedure:

Initialize the selected feature subset FS as an empty set.

For each feature ranking algorithm k in FRA:

a. Apply algorithm k to the dataset X to generate a ranking of features and their associated weights W_k.

For each weight vector W_i obtained:

a. Use K-Means clustering (treating each weight as a one-dimensional data point) to form a partition P_i.

Apply the BestOfK (BoK) strategy to select the most optimal partition from all P_i.

From the selected optimal partition, identify the cluster C_B that contains the features with the highest weights.

Add all features from C_B to FS.

Return FS as the final selected subset of features.

3. **Datasets with high dimensions (above 1000 features):** For datasets with an extremely high number of features (e.g., Lymphoma-4026, Colon-2000, and Leukemia-7129),

Received: August 08 2025

International Journal of Applied Mathematics
Volume 38 No. 5, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

applying first variation of the FREC algorithm results in a substantial feature subset. Conversely, Variation 2, which involves iterative K-means clustering, significantly increases the algorithm's complexity due to the need for more repetitions. Therefore, for these high-dimensional datasets, a modified approach is adopted: after a single iteration, the top 1% of features from the best cluster (which contains the highest-ranked features) are selected, and any unnecessary features are subsequently removedA distributed fuzzy rough set (DFRS) approach was recently proposed by [26] in cloud computing to address the expansion of huge data. DFRS's primary concept is to divide massive data sets into smaller segments, each of which is tasked with processing the fuzzy rough set on a cloud node. On 15 datasets from the UCI repository, they applied DFRS. The performance of the DFRS algorithm is measured in terms of its space complexity and time complexity. The FREC algorithm can be easily scaled on high dimensional datasets.

Table 1: UCI Benchmark datasets

Dataset	No.of Instances	No.of Features	No. of Classes
Lungcancer	32	56	2
Zoo	101	16	7
Wine	178	13	3
Ionophere	352	34	2
Musk1	476	168	2
Sonar	500	60	2
PIMA	768	8	2
Spambase	4601	57	2
Waveform	5000	40	3
Musk2	6598	168	2

Table 2: High-dimensional microarray data sets.

Dataset	No.of Instances	No.of Features	No. of Classes
Colon	2000	62	2
Leukemi	7129	72	2

Received: August 08 2025

International Journal of Applied Mathematics
Volume 38 No. 5, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

a			
Lympho ma	4026	66	3

5 Experimental Results and Analysis

5.1 Data sets used for FREC

Benchmark datasets from the UCI repository are used to test the suggested FREC algorithm. Among the data sets selected are the Musk1 and Musk2 that are medium-dimensional data sets, which contain 168 dimensions in total, and PIMA, that has a moderate amount of dimensions (8). Moreover, other high dimensional microarray gene expression cancer datasets are employed, including those for leukaemia (7129 features), lymphoma (4076 features), and colon (2000 features). The data sets used in this experiment are summarized in the Table 1 and Table 2.

5.2 FREC Implementation

Five different feature ranking algorithms are used from *Weka* to provide feature weights and various ranking criteria. These methods include Symmetric Uncertainty evaluators, ReliefF, OneR, InfoGain, and Gainratio. The characteristics are then divided into groups according to their weights using the K-means clus- tering technique. The ideal number of clusters is found using the Dunn Index [27]. To ascertain the ideal partitioning, BoK is utilised. The cluster containing all the features with the maximum weight is used to select the final feature subset. Using the chosen features from the data sets, a K-fold cross-validation is performed with K=10 for classification.

Figure 2 shows the process of FREC algorithm implementation on Wine data set.

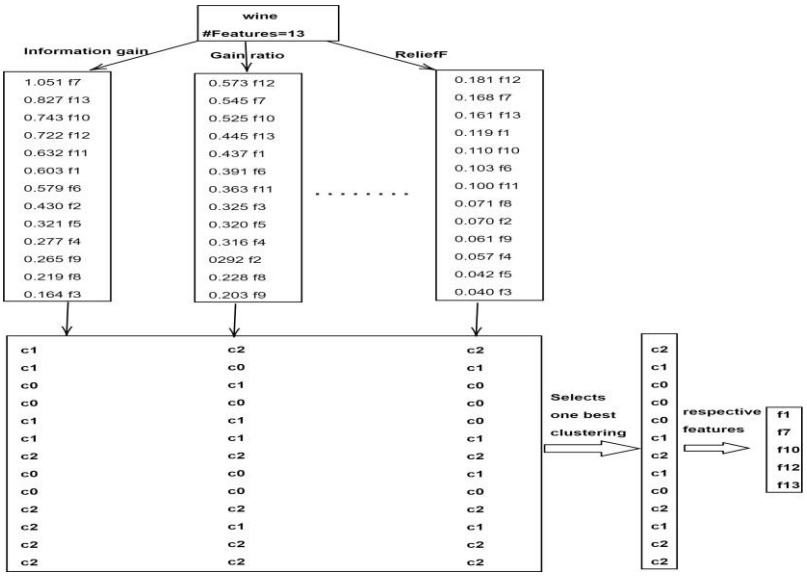

Fig. 2: Steps of FREC for Wine dataset.

Received: August 08 2025

International Journal of Applied Mathematics

Volume 38 No. 5, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

For datasets like Musk1 and 2 with medium number of features, an iterative application of the K-means clustering technique to the output of the preceding iteration progressively reduces the number of selected features. Classifier accuracy is evaluated after each iteration of this process. However, for high-dimensional microarray data sets, we choose only top 1% of features from the initial iteration's output due to considerable computational time required for repeated application.

Table 3: Comparison of features selected by FREC with literature

Dataset → Method↓	Wine(13)	Spambase(57)	Ionosphere(34)	Sonar(60)	Colon(2000)
FREC	**5**	**6**	**7**	**9**	**4**
GCNC [29]	8	26	15	26	43
GCACO [28]	7	25	18	25	45
RRFS [31]	9	33	22	34	56
UFSACO [30]	9	31	19	27	55

Table 4: Comparison of FREC classifier accuracy with literature

Dataset → Method↓	Wine(13)	Spambase(57)	Ionosphere(34)	Sonar(60)	Colon(2000)
FREC	97.19	86.10	92.30	75.40	87.02
GCNC [29]	96.18	82.11	88.91	77.36	85.47
GCA	96.73	86.02	91.14	72.50	78.14

Received: August 08 2025

International Journal of Applied Mathematics
Volume 38 No. 5, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

CO [28]					
RRFS [31]	95.42	84.71	90.40	73.53	79.96
UFSACO [30]	94.76	84.47	88.80	74.34	86.44

Additionally, duplicated features are eliminated from top 1% of the features, and the Random Forest method from *Weka* is used to assess the classifier's accuracy.

5.3 Results of FREC on UCI datasets

Tables 3 and 4 present the FREC method's performance in comparison to a few current approaches that are accessible in the recent literature.

As demonstrated in Table 3 and Table 4, FREC outperforms GCACO [28], GCNC [29], UFSACO [30], and RRFS [31] in terms of both accuracy and feature count in the context of Wine and Ionosphere datasets. Other approaches choose more than twice as many features as FREC, with the exception of Zoo and Wine data sets. The accuracies for the Spambase and Sonar datasets are 2% lower than those of the other methods under comparison since FREC selected a significantly smaller number of characteristics.

The results are tabulated in Table 3 and 4.

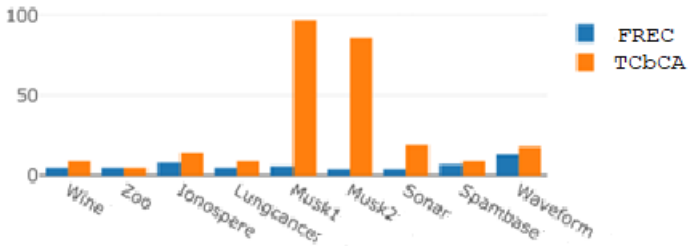

Fig. 3: Feature Selection by FREC compared to TCbGA

5.3.1 Comparative analysis of FREC and TCbGA

Table 5 details a performance comparison between FREC and TCbGA, a notable genetic algorithm-based method. This comparison is presented in a dedicated table because results for

Received: August 08 2025

International Journal of Applied Mathematics

Volume 38 No. 5, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

all the datasets we selected are available for both algorithms.

Table 5: FREC results compared to latest method TCbGA

Dataset	**FREC**		**TCbGA**[32]	
	#Features	Accuracy%	#Features	Accuracy%
Zoo(16)	**4**	96.88	5	98.03
Wine(13)	**4**	98.19	9	99.60
Ionosphere(34)	**6**	91.31	14	98.32
Waveform(40)	**11**	82.70	18	85.43
Lungcancer(56)	**7**	82.20	9	96.30
Spambase(57)	**6**	88.16	19	91.85
Sonar(60)	**8**	77.30	9	84.62
Musk1(166)	**10**	81.86	97	94.27
Musk2(166)	**15**	96.45	86	99.23

We compare our FREC method performance with the recent TCbGA algorithm. As indicated in Table 5, FREC consistently chooses a substantially smaller number of features than TCBGA. The feature count chosen by FREC is substantially fewer than the TCbGA[32] for low dimensional datasets such as Wine, Ionosphere, Lung can- cer, Spambase, Sonar, and medium dimensional datasets like Musk1.

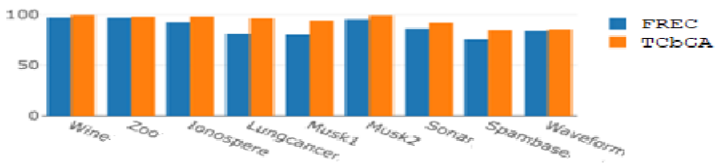

Fig. 4: FREC Classifier accuracy compared to latest TCbGA

Even fewer than 10% of Musk2's features were chosen using FREC. Although Table 5

Received: August 08 2025

International Journal of Applied Mathematics
Volume 38 No. 5, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

indicates that TCbGA's classifier accuracy is marginally greater than FREC's, FREC's feature selection is significantly more limited. It should be mentioned that TCbGA's computational complexity is far greater than ours. For instance, TCbGA took hours to find the best subset of features of the Sonar data set with a medium number of features (60) that are less than 100, but FREC took minutes to identify the optimal feature subset of any high-dimensional dataset.

5.3.2 Comparative analysis of FREC with classical methods

In Table 6 a comparison of FREC's performance against several conventional feature subset selection algorithms is given. When we compare the number of selected features with the literature, the algorithm features count is typically lower, with the exception of lung cancer, and the classifier accuracies are comparable. Additionally, the suggested method's computing complexity is significantly reduced. In most datasets, the number of characteristics chosen is reduced by roughly 50.

5.3.3 Microarray data sets results using FREC method

The most recent literature contains results for high dimensional microarray datasets. The results are reported in Table 7, which compares the performance

Table 6: Results of FREC compared to classical feature selection methods

Dataset	FREC		Literature	
	No. of Features	% of Accuracy	No. of Features	% of Accuracy
Pima(8)	**4**	**79.80**	5	79(SBC[17])
Zoo(16)	**5**	**96.20**	7	96(GARIPPER[33])
Wine(13)	**6**	**97.50**	7	97.40(GARIPPER[33])
Lung cancer(56)	**4**	**87.60**	5	87.00(ReliefF[34])
Ionosphere(34)	7	**95.30**	10	94.60(GARIPPER[33])
Waveform(40)	**10**	**82.70**	13	81.52(NMIFS[14])
Spambase(57)	**2**	**84.20**	3	75.8(NMIFS[14])
Sonar(60)	**8**	**88.40**	11	86.36(NMIFS[14])
Musk1(168	**5**	**76.30**	25	74.00(WBFS[

Received: August 08 2025

International Journal of Applied Mathematics

Volume 38 No. 5, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

)				35])
Musk2(168)	**6**	**95.60**	2 2	91.33(FCBF[16]) 94.6(ReliefF[34])

of FREC with some of these more recent techniques. Additionally, Tables 5 and 6 plot the results.

The results make it abundantly evident that the suggested FREC algorithm performs better on micro array datasets than any other algorithm. For the Leukaemia, Lymphoma, and Colon datasets, the FREC classifier's accuracy is higher than that of the literature; for the Lymphoma dataset, it is almost iden- tical to that of EnsRank. Nonetheless, the FREC's selection of characteristics for lymphoma and colon cancer is minuscule in comparison to the literature. FREC achieved similar accuracy as the Ensemble-Ranking (EnsRank) strategy [1], that selects 80 features, on the Lymphoma dataset with just 12 features. When FREC is compared to the FDT [37] approach, the FREC method chooses 26 characteristics in the leukaemia dataset. In summary, the classifier accuracy is higher than the literature, and the count of the features reduced to less than 0.5% of the total feature set. Comparing the computational complexity of FDT and FREC, the former is far more complex. In summary, the final selected features represent less than 1 percent of the overall features, which is a very tiny percentage compared to literature. The accuracy has been tested using the kNN classifier.

Table 7: Results obtained by on high dimensional datasets compared to the literature.

Data set	**FREC**		**Methods in Literature**	
	#Features	Acc%	#Features	Acc%
Colon	6	87.80	6	80.20(FDT [37])
Lymphoma	13	97.68	80	97.20 (EnsRank[1])
Leukemia	25	95.83	4	87.50(FDT [37])

On microarray data sets, the FREC algorithm performs better than alternative techniques

Received: August 08 2025

International Journal of Applied Mathematics
Volume 38 No. 5, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

in terms of competitiveness. Most of the approaches

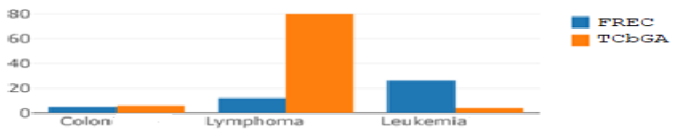

Fig. 5: Comparative Analysis of Feature Reduction in Microarray Datasets: FREC vs. Literature

mentioned in the literature require a specified threshold value to pick the final optimal subset of features, and also the value of threshold changes according to the total count of features. Anyhow, a predetermined threshold value is not necessary for FREC. Thresholds are only used when working with very large datasets. The findings unequivocally show that the features chosen using the FREC method are considerably better than those in the literature.

5.3.4 Robustness of FREC

We implement FREC using two new, recent Ensemble algorithms, WHAC and LWEA , presented by Huang et al., [38], and Banerjee et al., [39], respectively, in order to confirm the algorithm's resilience with various Ensemble clustering techniques.

Table 8 presents the results.The outcomes demonstrate that FREC is resilient to the Ensemble clustering algorithm selection.

Additionally, we have confirmed that the findings of FREC alone are sig- nificantly superior to the most recent approach, DFRS, when applied to the datasets listed in Table 9. Table 10 contains a tabulation of the results. The results unequivocally demonstrate that, in comparison to DFRS, FREC has achieved a large feature reduction for almost all data sets, along with a minor improvement in accuracy.

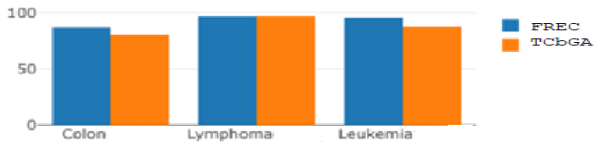

Fig. 6: Performance of FREC with literature

Received: August 08 2025

International Journal of Applied Mathematics

Volume 38 No. 5, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Table 8: Comparison of FREC with LWEA, BOK and WHAC methods

Dataset	**FREC with BoK**		**FREC+LWEA**		**FREC+WHAC**	
	#Features	**Acc%**	**#Features**	**Acc%**	**#Features**	**Acc%**
Wine(13)	4	96.19	5	98.13	5	98.13
Zoo(16)	4	98.02	4	97.22	4	97.22
Ionosphere(34)	7	94.30	6	94.12	6	94.12
Lung cancer(56)	4	84.10	6	82.10	6	82.10
Musk1(168)	7	81.46	5	81.10	4	80.20
Musk2(168)	5	96.30	5	94.30	5	94.30
Spambase(57)	5	88.10	5	85.10	5	85.10
Sonar(60)	8	75.40	5	76.40	5	76.40
Waveform(40)	12	84.80	9	82.53	9	82.53
Colon(2000)	6	86.02	4	86.02	5	87.62
Lymphoma(4096)	12	97.96	11	97.96	10	98.16
Leukemia(7129)	25	96.83	27	96.83	27	96.83

6 Conclusions

A novel approach is put forth to address the issues of large datasets and high dimensionality. The feature selection technique FREC performs effectively on datasets with up to 5000 occurrences and 7129 features. The FREC method falls under the category of heterogeneous ensemble approaches, which include using several algorithms on the same data. Compared to many contemporary approaches in the literature, FREC selects fewer features, and for the majority of datasets, accuracy is high. On high-dimensional

Received: August 08 2025

Volume 38 No. 5, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

microarray datasets, it is providing noticeably greater accuracy and feature reduction.

FREC is a greedy method that maintains a high classification accuracy while achieving a good feature reduction.

Table 9: UCI bigdata datasets

S.No	Dataset	No. of Classes	No.of Instances	No.of Features
1	Waveform	3	5000	21
2	Diagnosis	11	58509	49
3	HAPT	12	10929	561

Table 10: Results obtained by FREC recent methodDFRS

Dataset	FREC		DFRS	
	No.of Features	Accuracy%	No.of Features	Accuracy%
Waveform(21)	13	79.94%	17	77.78%
HAPT(561)	65	96.88%	347	94.49%
Diagnosis(49)	22	99.96%	26	96.23%

References

[1] Sandhya Rani, D., Sobha Rani, T., Durga Bhavani, S.: Feature subset selection using Ensemble clustering, 1–6 (2015)

[2] Rani, D.S., Rani, T.S., Bhavani, S.D.: Ensemble clustering for dimen- sionality reduction, 148–153 (2014). https://doi.org/10.1109/IC3.2014.6897164

[3] Swamy Das, M., Sandhya Rani, D., Reddy, C.R.K.: Heuristic based script identification from multilingual text documents, 487–492 (2012). https://doi.org/10.1109/RAIT.2012.6194627

Received: August 08 2025

Volume 38 No. 5, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[4] Swamy Das, M., Sandhya Rani, D., Reddy, C.R.K.: Feature Ranking Based Ensemble Clustering for Feature Subset Selection, pp. 1–6 (2024). https://doi.org/10.1007/s10489-024-05566-z

[5] Sandhya Rani, D., Sobha Rani, T., Durga Bhavani, S.: Graph-based feature selection using ensemble clustering. In: 2023 12th International Conference on Advances in Computing, Communication, and Control (ICAC3), pp. 1–6 (2023). https://doi.org/10.1109/ICAC3.2023.1023214

[6] Swamy Das, M., Sandhya Rani, D., Reddy, C.R.K.: Cascaded Two-Stage Feature Clustering and Selection Via Separability and Consistency in Fuzzy Decision Systems, pp. 1–6 (2024). https://doi.org/10.1109/ICAC3. 2024.1026783

[7] Gionis, A., Mannila, H., Tsaparas, P.: Clusetering aggregation. In: Pro- ceedings of International Conference on Data Engineering, ACM, pp. 341–352 (2005)

[8] Liu, H., Shao, M., Fu, Y.: Ensemble guided unservised feature selection. Proceedings of the Thirtieth AAAI Conference on Artificial Intelle- gence(AAAI16), ACM, 1874–1880 (2016)

[9] Liu, H., Wu, J., Liu, T., Tao, D., Fu, Y.: Spectral ensemble clustering via weighted kmeans: Theoretical and practical evidence. IEEE Transactions on Knowledge and Data Engineering, 1129–1143 (2017)

[10] Liu, H., Setiono, R.: Chi2: Feature selection and discretization of numeric attributes. Proceedings of the Seventh International Conference on Tools with Artificial Intelligence, IEEE, 388–391 (1995)

[11] Hall, M.A., Holmes, G.: Benchmarking attribute selection techniques for discrete class data mining. IEEE Transactions on Knowledge and Data Engineering, 15(6), 1437–1447 (2003)

[12] Kenji, K., Larry A, R.: The feature selection problem: Traditional methods and a new algorithm. In: Proceedings of AAAI, pp. 129–134 (1992)

[13] Yu, L., Liu, H.: Efficient feature selection via analysis of relevance and redundancy. Journal of Machine Learning Research **4**, 1205–1224 (2004)

[14] Estevez, P.A., Tesmer, M., Perez, C.A., Zurada, J.M.: Normalized mutual information feature selection. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, pp. 189–201

[15] Peng, H., Long, F., Ding, C.: Feature selection based on mutual informa- tion: Criteria of maxdependency, maxrelevance, and minredundancy. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27,pp. 1226–1238 (2005)

[16] Yu, L., Liu, H.: Feature selection for highdimensional data:a fast correla- tionbased filter solution. In: Proceedings of the Twentieth International Conference on Machine Learning (ICML), pp. 856–863 (2003)

Received: August 08 2025

Volume 38 No. 5, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[17] Ratanamahatana, C.A., Gunopulos, D.: Feature selection for the naive bayesian classifier using decision trees **17**, 475–488 (2006)

[18] Song, Q., Ni, J., Wang, G.: A fast clustering based feature subset selection algorithm for high dimensional data. In: IEEE Transactions on Knowledge and Data Engineering, vol. 25, pp. 1–14 (2013)

[19] Press, W.H., Flannery, B.P., Teukolsky, S.A., Vetterling, W.T.: Numerical recipes in c. Cambridge University Press, Cambridge (1988)

[20] Zhang, Z., Hancock, E.R.: Hypergraph based information theoretic feature selection. Pattern Recognition Letters **33**, 1991–1999 (2012)

[21] Monalisa, M., Mukhopadhyay, A.: Unsupervised nonredundant featureselection: a graphtheoretic approach. In: Proceedings of the Interna- tional Conference on Frontiers of Intelligent Computing: Theory and Applications(FICTA), pp. 373–380 (2013)

[22] Bandyopadhyay, S., Bhadra, T., Mktra, P., Maulik, U.: Integration of dense subgraph finding with feature clustering for feature selection, vol. 40, pp. 104–112 (2014)

[23] Hong, Y., Kwong, S., Chang, Y., Ren, Q.: Ensemble unsupervised feature ranking from multiple views. Pattern Recognition Letters,Elsevier **29**(5), 595–602 (2008)

[24] Jong, K., Mary, J., Cornuejols, A., Marchiori, E., Sebag, M.: Ensem- ble feature ranking. Knowledge discovery in databases:PKDD, 267–278 (2004)

[25] Sandhya Rani, D., Sobha Rani, T., Durga Bhavani, S.: Feature subset selection using Ensemble clustering. In: 2015 Eighth International Con- ference on Advances in Pattern Recognition (ICAPR), pp. 1–6 (2015). https://doi.org/10.1109/ICAPR.2015.7050659

[26] Kong, L., Qu, W., Yu, J., Zuo, H., Chen, G., Xiong, F., Pan, S., Lin, S., Qiu, M.: Distributed feature selection for big data using fuzzy rough sets. IEEE Transactions on Fuzzy Systems **PP**, 846–857 (2019). https://doi.org/101109/TFUZZ20192955894

[27] Kovacs, Legany, F., Babos, A.: Cluster validity measurement tech- niques. In: Proceedings of the 6th International Symposium of Hungarian Researchers on Computational Intelligence, pp. 18–19 (2005)

[28] Moradi, P., Rostami, M.: Integration of graph clustering with ant colony optimization for feature selection. Knowledge based systems **84**, 144–161 (2015)

[29] Moradi, P., Rostami, M.: A graph theoretic approach for unsupervised feature selection. Engineering Applications of Artificial Intelligence **44**, 33–55 (2015)

[30] Tabakhi, S., Moradi, P., Akhlaghian, F.: An unsupervised approach to feature selection algorithm based on ant colony optimization. Engineering Applications of

Received: August 08 2025

Volume 38 No. 5, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Artificial Intelligence **32**, 112–123 (2014)

[31] Ferreira, A.J., Figueiredo, M.A.T.: An unsupervised approach to feature descritization and selection. Pattern recognition **45**, 3048–3060 (2012)

[32] Ma, B., Xia, Y.: A tribe competitionbased genetic algorithm for feature selection in pattern classification. Applied Soft Computing **58**, 328–338(2017)

[33] Yang, J., Tiyyagura, A., Chen, F., Hanover, V.: Feature subset selec- tion for rule induction using ripper. In: Proceedings of Genetic and Evolutionary Programming, pp. 117–136 (1998)

[34] Megchelenbrink, W., Marchiori, E., Lucas, P.: Relief based feature selec- tion in bioinformatics:detecting functional specificity residues from mul- tiple sequence alignments. Master Thesis, Radboud University, Nijmegen (2010)

[35] Leng, J., Valli, C., Armstong, L.: A wrapper based feature selection for analysis of large data sets. In: Proceedings of 3rd International Confer- ence on Computer and Electrical Engineering(ICCEE), IEEE Computer Society, pp. 167–170 (2010)

[36] Hall, M.: Correlationbased feature selection for machine learning. PhD thesis, Citeseer (1999)

[37] Ludwig, S.A., Picek, S., Jakobovic, D.: Chapter 13: Classification of cancer data: Analyzing gene expression data using a fuzzy decision tree algo- rithm. In: Operations Research Applications in Health Care Management, International Series in Operations Research & Management Science 262, Springer (2018)

[38] Huang, D., Wang, C., Lai, J.: Locally weighted ensemble clustering. IEEE Transactions on Cybernetics **48**, 1460–1473 (2018). https://doi.org/ 101109/TCYB20172702343

[39] Banerjee, A., Pujari, A.K., Panigrahi, C.R., Pati, B., Nayak, S.C., Weng, T.: A new method for weighted ensemble clustering and coupled ensemble selection. Connection Science, Taylor & Fran- cis **33**(3), 623–644 (2021) https://arxiv.org/abs/https://doiorg/101080/0954009120201866496. https://doi.org/101080/0954009120201866496

[40] D., Sandhya Rani, Rani, T. S., Bhavani, S. D., & Krishna, G. B. (2024). Feature ranking based Ensemble clustering for feature subset selection. *Applied Intelligence*, *54*(7), 8154–8169. https://doi.org/10.1007/s10489-024-05459-7

Received: August 08 2025