Journal of Harbin Engineering University
ISSN: 1006-7043
Vol 44 No. 7
July 2023

Cluster Validity Indices for Detect the Optimal Clusters Feature Sets in Directive Elbow and Cluster-Wise Feature Selection

Suman Laha [1,2], Utpal Roy [1], Amit Kumar Saxena [2,*], Damodar Patel [2], Rajeshwar Prasad [2]
[1] Department of Computer & System Sciences, Visva-Bharati University, Santiniketan, WB, 731235, India, mrlaha@gmail.com, utpal.roy@visva-bharati.ac.in
[2] Department of Computer Science & Information Technology, Guru Ghasidas Vishwavidyalaya, Bilaspur, CG, 495009, India, mrlaha@gmail.com, amitsaxena65@rediffmail.com, damodarpatel7497@gmail.com, rp4464867@gmail.com

Abstract-In this empirical model, we first detect the optimal number of clusters in unsupervised feature set using three different coefficient (CVI) measures computed under varying *K* of K-means. To determine learned optimal number of clusters in the feature set, we use directive elbow method i.e., a graphical solution which provides the optimal point in the experimental data of *K* versus coefficient value. Three Cluster Validity Indices (CVI) such as Distortion (DIS), Inertia (INE) & Silhouette (SIL) are being used in directive elbow method; each of them is providing respective optimal number of clusters in a feature set.
Once we have the learned optimal number of clusters in a feature set, we select the most significant feature from each cluster of features using an unsupervised feature selection method which is based on locality preserving power in terms of Laplacian Score [1] of feature.
To measure the efficiency of this model, we look for the learned optimal number of clusters in data points using directive elbow method. New class labels are provided by the K-means using learned optimal number of clusters in data points. Initial class labels (found with dataset source) were preserved. We compare classification accuracies between all features and selected features based on initial class labels and new class labels using two renowned machine learning techniques KNN & SVM classifiers.
This empirical model selects non redundant and most relevant features in unsupervised scenario. This model does not require any predefined assumption of the values of parameters such as number of cluster and number of selected features. The model has been tested on medium to high dimensional (9 to 11,000 features) twelve heterogeneous data sets and found the improvement (or no significant drop) in classification accuracy obtained on selective features.

Keyword-machine learning, unsupervised feature selection (UFS), feature clustering, K-means clustering, Laplacian score, dimensionality reduction, Cluster Validity Indices (CVI), distortion index, inertia index, Silhouette index, elbow method

1. Introduction & Related Work

High dimensional data is prone to bulky redundant features. Due to redundant features, we face performance issues in machine learning models and this phenomenon is called curse of dimensionality[2]. Due to these redundant features; huge amount of unnecessary memory is consumed during the processing of machine learning algorithms; computational burdens are increased and visual presentation of the data becomes either very cumbersome or impossible task. Irrelevant or noisy features are also responsible for the degradation in performance of machine learning models.

Among all dimensionality reduction techniques, Feature Selection (FS) is a popular and emerging research topic in machine learning. FS involves two important tasks such as removal of identical or redundant features and selection of most relevant features which could retain most of the knowledge found in dataset.

Feature selection algorithms can be divided into two categories based on the feature evaluation methodology; wrapper approach [3] and filter

approach. Generally, wrapper approaches [4] use a particular classifier as the measure of importance (significance) for a candidate feature subset. Firstly, the wrapper produces a candidate feature subset by the search strategy, and then the classifier is trained and tested to evaluate the candidate feature subset. This process will be iteratively performed until the selected feature subset meets the specific requirements. Filter approaches attempt to compute the importance of the individual features and assign each feature with a score without using any classifier. The set of features could be sorted with respect to their scores. Then important features are selected and forwarded to the classifier whereas the less important features (having comparative poor score) are discarded.

With respect to the availability of class labels, FS methods could be divided in two major groups called supervised and unsupervised. The association between features and class labels is a crucial fact while selecting features in any supervised FS and the selected features should have a strong relationship to class labels. In case of high dimensional modern dataset, absence or inadequacy of class label promotes unsupervised feature selection (UFS).

Cluster analysis in a multidimensional space is to find a set of groups or clusters where each group comprises similar objects which are more similar to each and other, than the objects belong to other groups. Criterion of effective cluster outcome is that, intra cluster (group) similarity and inter cluster dissimilarity both should be considerably high. Evaluation of this criterion could be performed using different Cluster Validity Indices (CVI) [5]–[8]. In some CVI such as distortion and inertia, graphical plot of varying cluster number (x-axis) versus corresponding CVI (y-axis) gives heuristic solution of appropriate number of clusters. Flattened plots in the graph indicate that, respective clusters are sufficient dissimilar. The search for an Elbow [9] point in the graph provides the appropriate number of clusters.

To get rid of the feature redundancy problem, some existing research papers [10]–[13] proposed UFS methods built on feature cluster algorithms. In these papers, renowned clustering algorithms have been used to perform the feature cluster analysis and they were able to minimize the redundancy among the feature set. Clustering algorithms have some innate shortcomings; one of them is supply of assumed values of parameters at the beginning of processing and appropriate speculation about these values of parameters plays a vital role in feature selection process. Model performance is typically dependent on the assumption of values of these parameters.

Alfian G et al. [14] in his paper used extra-trees classifiers to remove irrelevant features and then prediction model applied on breast cancer data. Zivkovic M et al. [15] forwarded swarm intelligence-based method for extracting the most relevant features from datasets. Novel Meta-Heuristic Algorithm for feature selection was suggested by El-Kenawy et al. [16] by amalgamating Sine Cosine Hybrid Optimization with Modified Whale Optimization for solving binary decision variables. Abdollahzadeh B et al. [17] amalgamated two different algorithms called Harris Hawks Optimization (HHO), Fruitfly Optimization Algorithm (FOA) into a third modified one called MOHHOFOA which is used to execute the feature selection. Relevance based on Weight (RW) are two changed ratios which could predict feature relevance and used in multi-level feature selection proposed by Hu L et al. [18]. Tiwari A et al. [19] addressed metaheuristic & information-theoretic based problems for constructing feature selection algorithms and proposed an iterative feature selection hybrid method using Dynamic Butterfly Optimization Algorithm based Interaction Maximization and mutual information-based Feature Interaction Maximization (FIM). Nadimi-Shahraki M et al. [20] proposed an enhanced whale optimization algorithm (E-WOA) to strengthen WOA's low population diversity & way weak search strategy in NP-hard problem like feature subset selection. Zhu P et al. [21] enhanced the performance of Graph-based unsupervised feature selection

methods by integrating similarity matrix construction and feature selection process together as a simultaneous process in an algorithm named Graph Learning Unsupervised Feature Selection (GLUFS). An algorithm that relies on boosting/ sample re-weighting to subset good features, derived from tree-boosting models is proposed by Alsahaf A et al. [22]. Wahid A et al. [23] addressed the issues of outliers and noise in unsupervised learning and proposed Unsupervised Feature Selection with Robust Data Reconstruction (UFS-RDR) in which Mahalanobis distance is used for the detection of outliers & the graph regularized weighted data reconstruction error function is minimized. Huang P et al. [24] proposed unsupervised feature selection algorithm Adaptive Graph and Dependency Score (AGDS) that combines two methods i.e., adaptive-graph-based, and self-representation-based. Ouadfel S et al. [25] proposed a hybrid feature selection approach called RBEO-LS based on the ReliefF filter method and a novel metaheuristic Equilibrium Optimizer (EO) which is applied on high dimensional datasets. To rank features as per their relevancy Laplacian scores are used and using a wrapper approach the best subset of ranked features is chosen in a proposed scheme of paper of Patel D et al. [26]. Wu J et al. [27] proposed a novel unsupervised feature selection framework effective for high dimensional data, that they called as dictionary learning by eliminating redundancy and noise using dual sparse regression. To partition lung tissues and regions of Interest (ROIs) from the CT slices feature selection is performed first through a wrapper technique using manta ray foraging optimization (MRFO) algorithm and random forest (RF) classifier in the paper of Isaac A et al. [28]. To overcome difficulties in achieving good convergence that leads to increased computation in high dimensional feature selection using genetic Algorithm (GA), Abdulhussien A et al. [29] introduced multifeature fusion and discriminant FS using a novel quantum inspired GA (QIGA) to be essentially used in a good offline signature verification (OSV) system to avoid fraud. To get feature subset having optimal cardinality Agrawal S et al. [30] introduced an algorithm for multimodal multiobjective optimization based on differential evolution. Zivkovic M et al. [31] proposed a method by combining machine learning with the metaheuristic approaches (learnheuristics) through the implementation of an algorithm which is a modified version of the salp swarm algorithm for feature selection. Saxena A et al. [32] proposed a Genetic Algorithm based model where Sammon Error used as fitness value and the iterations done on randomly generated small partitions of features to get desired small number of features. Chugh D et al. [33] proposed a new unsupervised feature selection method named as densest feature graph augmentation by finding the maximally non-redundant feature subset in the first phase and then disjoint features are added to the feature set in the second phase. Patel D et al. [34], in their paper, propose a novel algorithm to select small & best subset of features using different combinations of feature across a number of trials through a wrapper based technique by which they improve the accuracy and the cardinality of the selected feature set.

In our paper, we propose a directive process which could compute the optimum number of clusters in the set of feature points. Directive elbow method finds optimum number of clusters in feature points. It uses three Cluster Validity Indices (CVI) such as Distortion (DIS), Inertia (INE) & Silhouette (SIL) under varying K of K-means. It provides the optimum cluster validity in order to obtain optimum value of K in each CVI. Features within the same cluster are redundant as they are similar and features picked up from different clusters are dissimilar. We select the best feature from each cluster and hence get K number of selected features. Selection of the best feature of a cluster is done by picking up the most relevant feature based on locality preserving power in terms of Laplacian score of features within the cluster. Therefore, this model is free from assumptions of parameters such as number of cluster and number of selected features.

2. Critical Details

2.1 Cluster Validity Indices (CVI)

Criterion of effective cluster outcome is that, intra cluster (group) similarity and inter cluster dissimilarity both should be considerably high. Evaluation of this criterion could be performed using different Cluster Validity Indices (CVI). In this paper, we use feature clustering as a tool to remove redundancy among features. We require appropriate clustering quality and picking up the most appropriate clustering is done by the directive Elbow (as described in 3.1). We use three renowned CVIs to measure the quality of clustering as described below:

2.1.1 Distortion (DIS)

In K-means, accumulation of squared distances between the cluster center and respective objects defines the distortion for a cluster which could be expressed as: $DIS_i = \sum_{j=1}^{N_i}[distance(x_{ij}, w_i)]^2$, where, DIS_i distortion for i^{th} cluster, N_i is number of objects in the i^{th} cluster, x_{ij} is the j^{th} object of the i^{th} cluster , w_i is center of the i^{th} cluster and $distance(x_{ij}, w_i)$ is the distance between x_{ij} and w_i. To get the distortion for all clusters, we accumulate all individual cluster distortions as: $DIS_K = \sum_{i=1}^{K} DIS_i$ where, K is the number of clusters. DIS_K denotes the overall distortion for the clustering having K number of clusters. Finally, the average distortion is:

$$DIS = \frac{1}{N} DIS_K$$

(1)

Where, N is the total number of objects considered for clustering.

2.1.2 Inertia (INE)

The squared distance from a clustered object to its closest centroid is the inertia of that object. After clustering, if we accumulate the inertia of each portioned object we get the inertia of the partitioned objects i.e.,

$$INE = \sum_{i=1}^{N} \min \left|\left|x_i - \mu_j\right|\right|^2 \mid \mu_j \in \mathrm{C}$$

(2)

Where, N is the total number of objects considered for clustering, μ_j is the closest centroid for x_i and C is the set of centroids.

2.1.3 Silhouette (SIL)

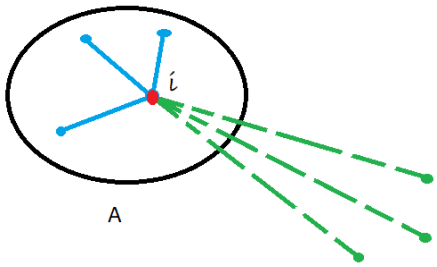


Figure-1: Silhouette of a data object

If we assume with an example, out of the clusters partitioned by any clustering technique, we draw one cluster A. Cluster A contains four objects; i in red dot and three other objects in blue dots. Three green dots are all remaining objects from clusters other than cluster A. Within cluster average dissimilarity of the object i; $a(i)$ is average of sum of distances (blue lines) measured from i to all other objects (blue dots) of A. Outside cluster average dissimilarity of the object i; $b(i)$ is minimum of average of sum of distances (green lines) measured from i to all objects (green dots). The Silhouette Cluster Validity Index for the particular object i could be written as: $s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}}$. The average $s(i)$ over all objects of a dataset measures the correctness of partitioning technique. Silhouette Cluster Validity Index for a dataset partitioned by any clustering technique having at least two disjoint clusters is written as:

$$SIL = \frac{1}{N}\sum_{i=1}^{N} s(i)$$

(3)

where, N is the total number of objects in the dataset. SIL compiles both compactness and severalty of the separated clusters in form of a ratio and its value of $s(i)$ ranges in between -1 to $+1$. $+1$ means all the clusters are extremely compact and severalty of clusters is at highest i.e., clusters are portioned most perfectly whereas

value near to -1 means clusters are not compact and no severalty of clusters is found i.e., objects are not partitioned properly.

2.2 Elbow

In this method, we graph the coefficient (CVI) on the y-axis and the number of clusters on the x-axis. The clusters are being combined by the ongoing clustering technique and the dissimilarity in clusters is being identified by the flattening of the graph. To find the appropriate number of clusters we mark the "elbow" point of the graph. We suggest Directive Elbow (3.1) method to determine the optimal point in the experimental data of number of clusters versus coefficient (CVI) value.

2.3 Laplacian Score (LS) [1]

Central goal of LS is to evaluate features according to their locality preserving power. The evaluation is typically dependent on the local structure of the data instead of the global structure of the data. To represent local structure of the data respective nearest neighbor graph is used. Every feature respect this graph structure to some extent. For a particular feature, the degree of respect defines its importance in the feature set. The algorithm to find the LS for a feature is defined here.

Let, we assume L_r is the LS of the r^{th} feature and f_{ri} is the i^{th} sample of the r^{th} feature and $i = 1, \dots, m$ where, m is the number of data points.

2.3.1. G is a nearest neighbor graph with m nodes to be constructed. x_i represents the i^{th} node. If x_i and x_j are "close" i.e. x_i is among k nearest neighbors of x_j or x_j is among k nearest neighbors of x_i; an edge between i^{th} and j^{th} node is considered i.e. nodes i and j are connected.

2.3.2. Express $S_{ij} = e^{-\frac{\|x_i - x_j\|^2}{t}}$, if nodes i and j are connected (here t is a suitable constant). Else, Express $S_{ij} = 0$. S represents the weight matrix of G which models the local structure of the data space.

2.3.3. For the r^{th} feature, we define:

$f_r = [f_{r1}, f_{r2}, \dots, f_{rm}]^T$, $D = diag(S1)$, $1 = [1, \dots, 1]^T$, $L = D - S$,

where L is a matrix called graph Laplacian [35]. Let

$$\tilde{f}_r = f_r - \frac{f_r^T D 1}{1^T D 1} 1$$

2.4.4. LS of the r^{th} feature is defined as follows:

$$L_r = \frac{\tilde{f}_r^T L \tilde{f}}{\tilde{f}_r^T D \tilde{f}} \qquad (4)$$

3. Proposed Methodology

3.1 Directive Elbow

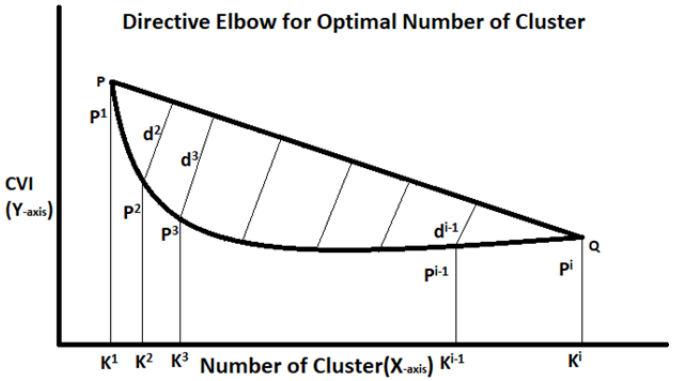


Figure-2: Directive Elbow for detecting optimal number of clusters

Considering i number of iterations, computations of Distortion & Inertia under varying K of K-means produce the experimental data like $\{K^i, Y^i\}$ where K^i represent number of clusters in i^{th} iteration and Y^i represent corresponding CVI value in i^{th} iteration. From this data, we plot the graph using set of points $\{P^i\}$ where P^i represents the data point found at i^{th} iteration. We find a straight line PQ using the terminals P^1 & P^i i.e., the first and last point. Now, we compute shortest distances $\{d^i\}$ from every point of $\{P^i\}$ to the straight line PQ. Therefore, the Elbow point represents the point having maximum$\{d^i\}$; directs the optimal number of clusters on x-axis.

For Silhouette, the Elbow point represents the point having maximum $\{Y^i\}$; directs the optimal number of clusters on x-axis.

3.2 Cluster wise Feature Selection

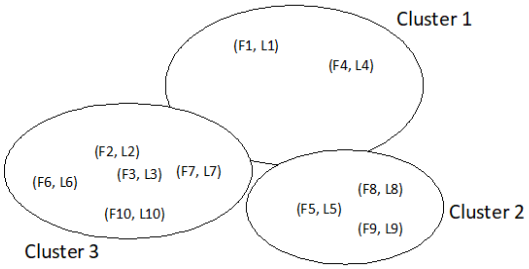


Figure-3: Cluster wise feature selection

If we assume with an example, ten features F1 to F10 and their LS score is labeled L1 to L10 respectively where L1 score has more relevancy with respect to L2 score and L2 score has more relevancy than L3 score and so on i.e., F1 is the most relevant feature and F10 is the least relevant feature in this example scenario. Optimized feature clustering provides three clusters cluster 1, cluster 2 and cluster 3. Cluster 1 contains {F1, F4}, Cluster 2 contains {F5, F8, F9} and Cluster 3 contains {F2, F3, F6, F7, F10}. Since, cluster analysis in a multidimensional space is to find a set of groups or clusters where each group comprises similar objects which are more similar to each and other, than the objects belong to other groups. Therefore, we chose one best feature from each cluster i.e., three features are to be selected from three clusters in compliance with the redundancy issue. From cluster 1 we select F1 as L1 score directs more relevancy in comparison with L4. From cluster 2 we select F2 as L2 score directs more relevancy in comparison with L3, L6, L7 and L10. From cluster 3 we select F5 as L5 score directs more relevancy in comparison with L8 and L9. Thus, we have selected features {F1, F2, F5} in this case.

3.3 Block diagram for Feature Selection

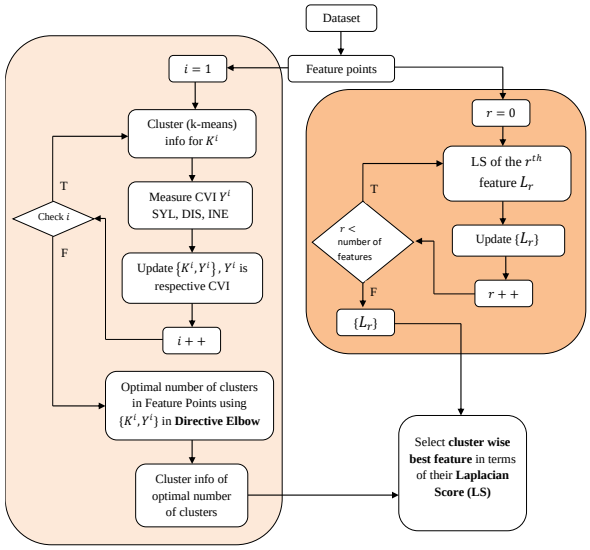


Figure-4: Block diagram

We cluster feature points using renowned clustering technique k-means iteratively using varying cluster numbers. Iterations one to at most fifteen ($i \leq 15$) supplies K^i and in each iteration three CVI methods SYL (as per Equation-3), DIS (as per Equation-1) and ITE (as per Equation-2) measurements are taken to update $\{K^i, Y^i\}$. Once updated $\{K^i, Y^i\}$ is achieved directive elbow (depicted in 3.1) finds the optimal number of clusters. Respective cluster information is forwarded for computation of the selected features. The number of selected features is set equal to the optimal number of clusters. Parallelly, we compute Laplacian score L_r (as per Equation 4) for each feature and $\{L_r\}$ is forwarded for computation of the selected features. Finally, we select cluster wise best feature in terms of their LS score (depicted in 3.2).

4. Experimental Results

4.1 Dataset used

Table-1: Datasets

Data	Instances	Features	Class
diabetes	520	16	2
arcene	200	10000	2

CLL_SUB	111	11340	3
HCV	615	12	5
BreastCancerWisconsin	699	9	2
madelon	2600	500	2
PCMAC	1943	3289	3
student	1044	33	4
colon	62	2000	2
orlraws10p	100	10305	10
sonar	208	60	2
librasMovement	360	90	15

This experiment used twelve datasets taken from UCI Machine Learning Repository and Feature Selection at Arizona State University [36], [37]. Table: 1 shows the description of datasets: name of the dataset, number of instances, number of features and number of classes. Datasets have wide range in numbers of (9 to 11,000) features. Datasets are chosen and preprocessed in such a way that there are no missing values and the last column is the class label, rest columns are attributes. The original sequence (serial numbers of columns index started with zero) of features of the datasets remains same as found in the repository.

4.2 Optimal number of clusters in feature points using Directive Elbow

We have plotted the learned optimal number of clusters in feature points in red dot as per the Decretive Elbow (described in 3.1). We have shown the result of Directive Elbow on DIS, ITE & SIL for each dataset (Figure-5 to 16).

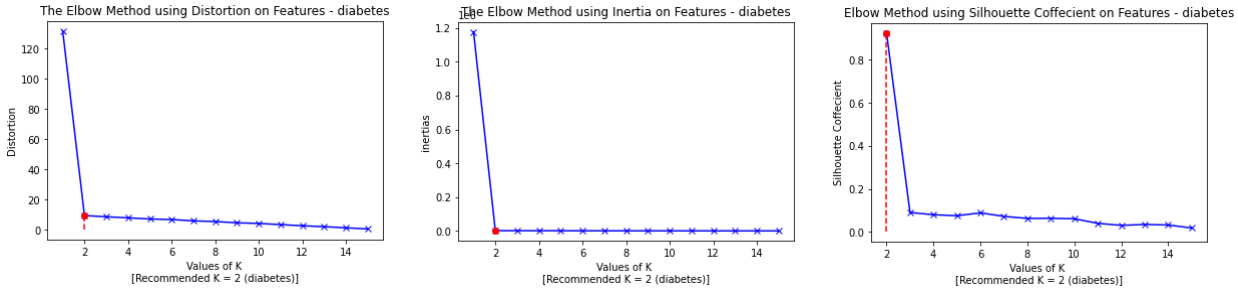


Figure-5: Optimal number of clusters in feature points – diabetes

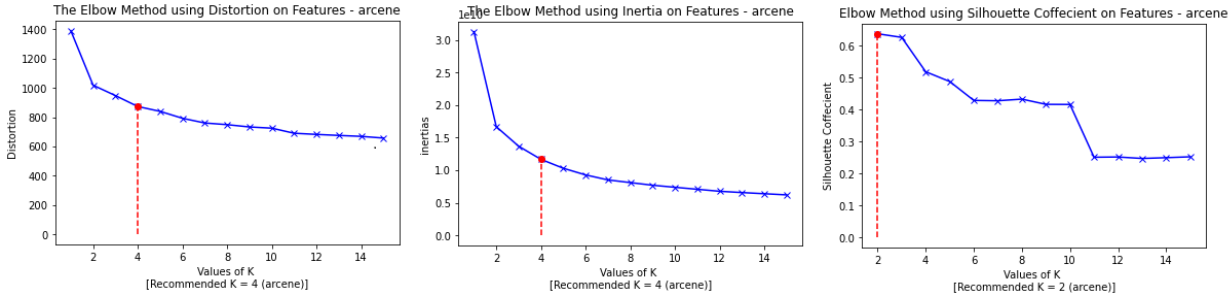


Figure-6: Optimal number of clusters in feature points – arcene

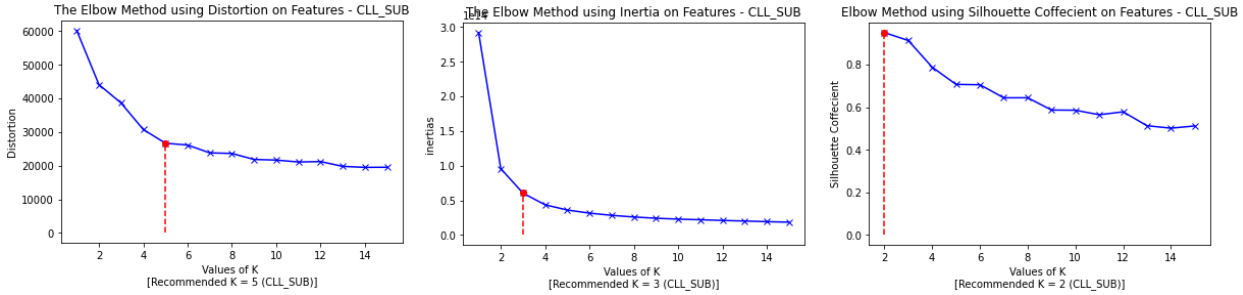


Figure-7: Optimal number of clusters in feature points – CLL_SUB

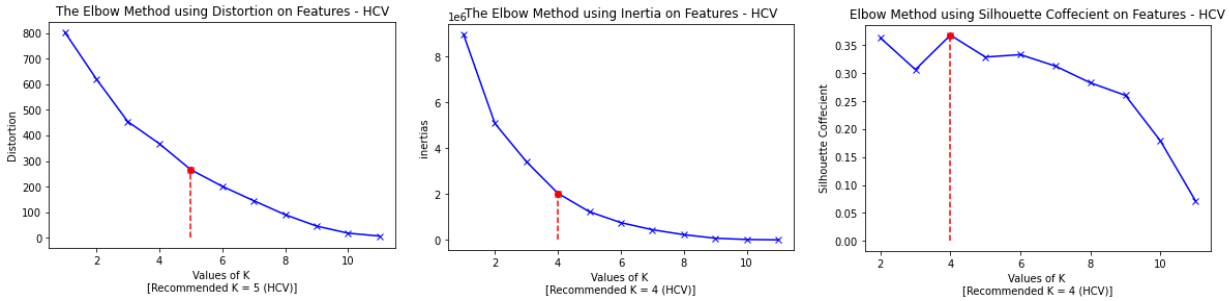


Figure-8: Optimal number of clusters in feature points – HCV

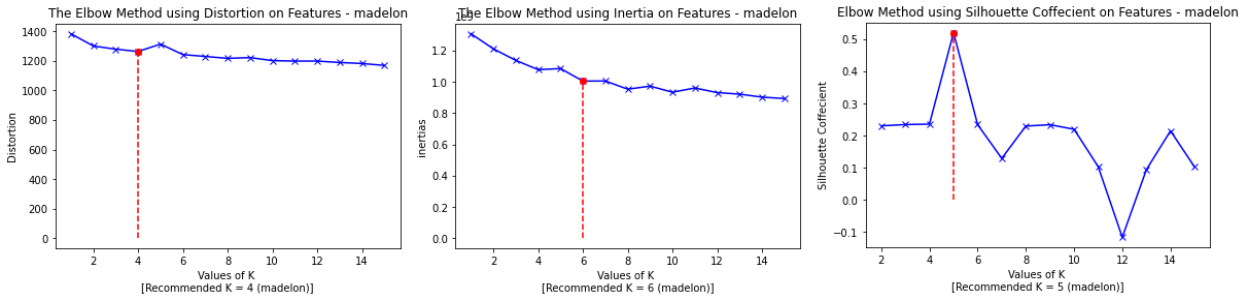


Figure-9: Optimal number of clusters in feature points – Breast Cancer Wisconsin

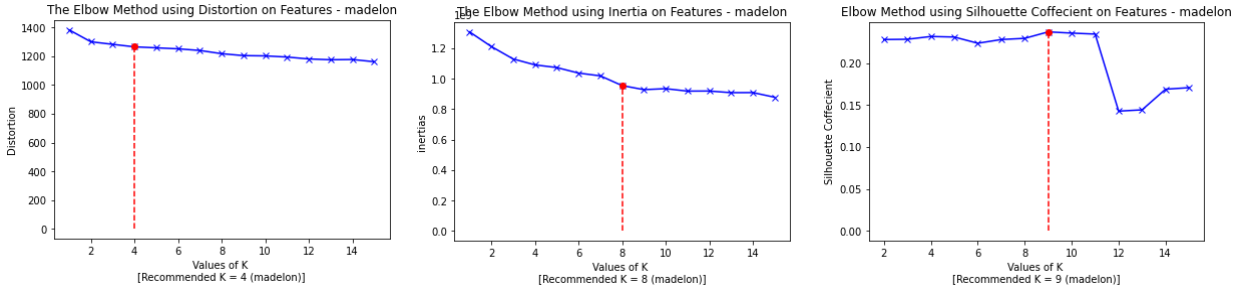


Figure-10: Optimal number of clusters in feature points – madelon

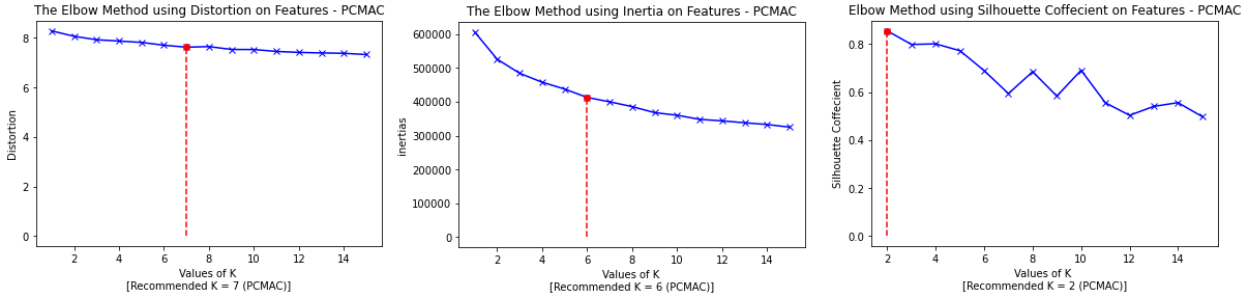


Figure-11: Optimal number of clusters in feature points – PCMAC

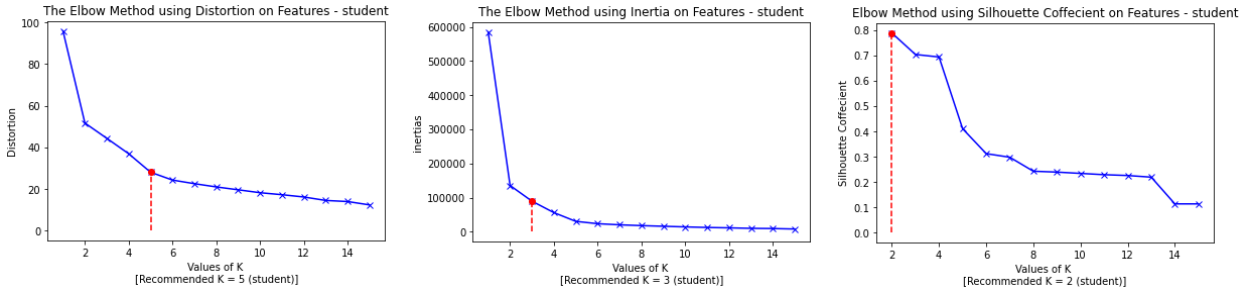


Figure-12: Optimal number of clusters in feature points – student

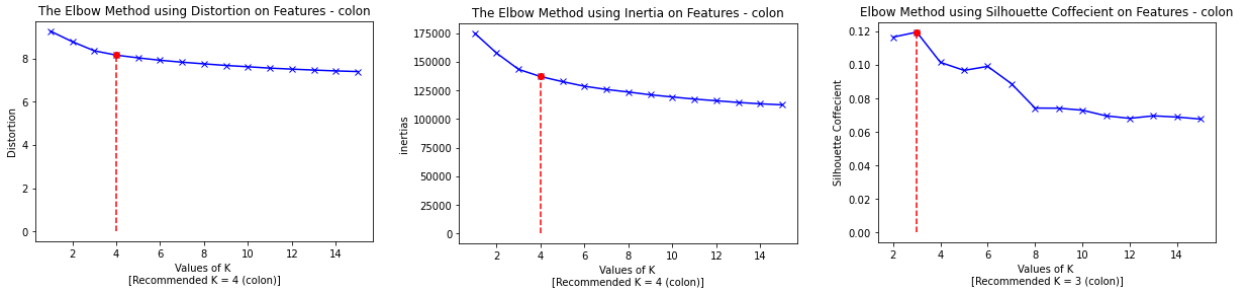


Figure-13: Optimal number of clusters in feature points – colon

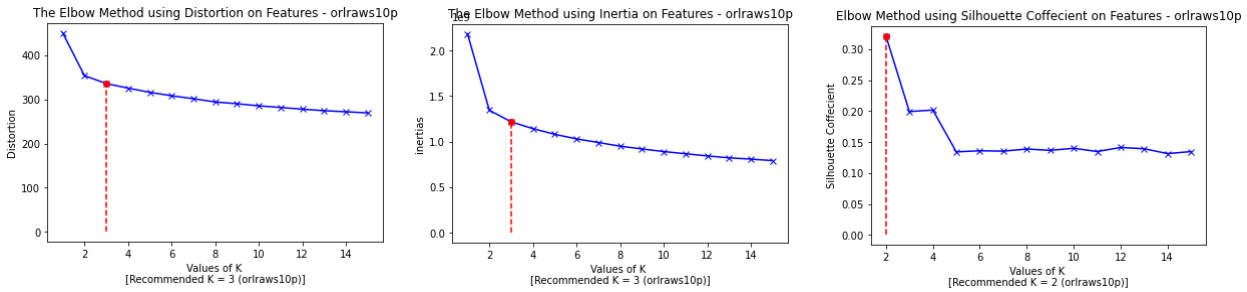


Figure-14: Optimal number of clusters in feature points – orlraws10p

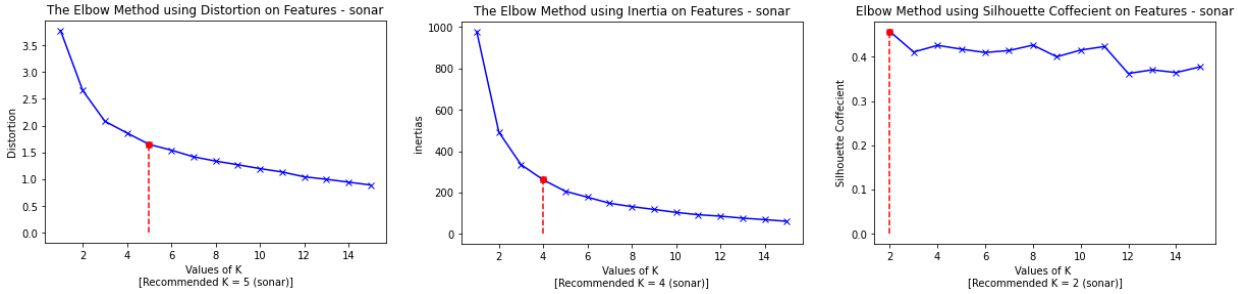


Figure-15: Optimal number of clusters in feature points – sonar

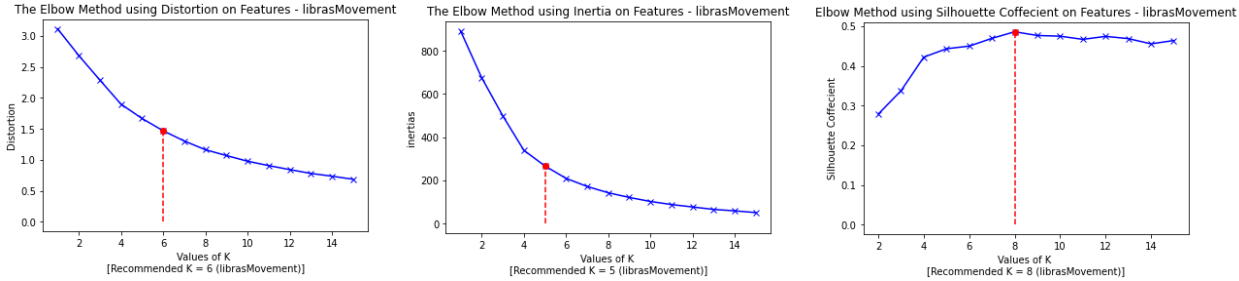


Figure-16: Optimal number of clusters in feature points – libras Movement

4.3 Feature redundancy

Redundancies of selected features are verified using correlation heatmap (Figure-17 to 28). We plotted feature to feature correlation matrix and almost all the cases verify our directive elbow selects optimal number of clusters in a feature set adequately good so that the feature clusters are compact and separated well. Thus, selected features are non redundant in basis of their correlation heat map. A contrary in the arcene data (Figure-18), inclusion of one high redundant feature observed when four features are selected.

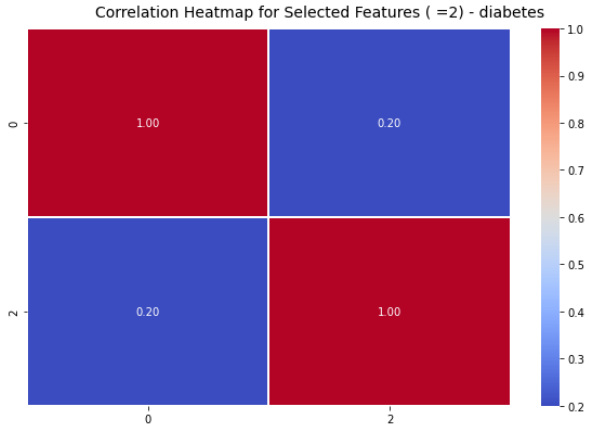


Figure-17: Correlation Heatmap for diabetes

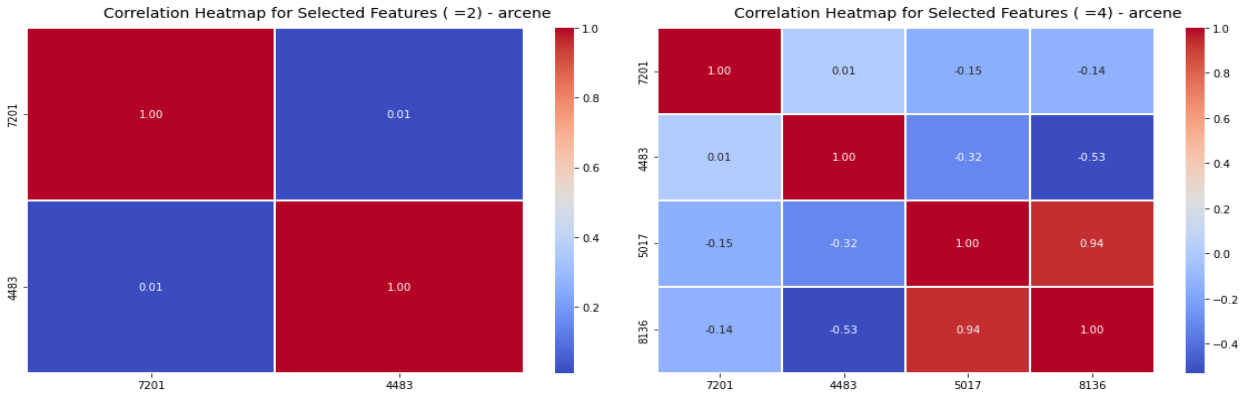


Figure-18: Correlation Heatmap for arcane

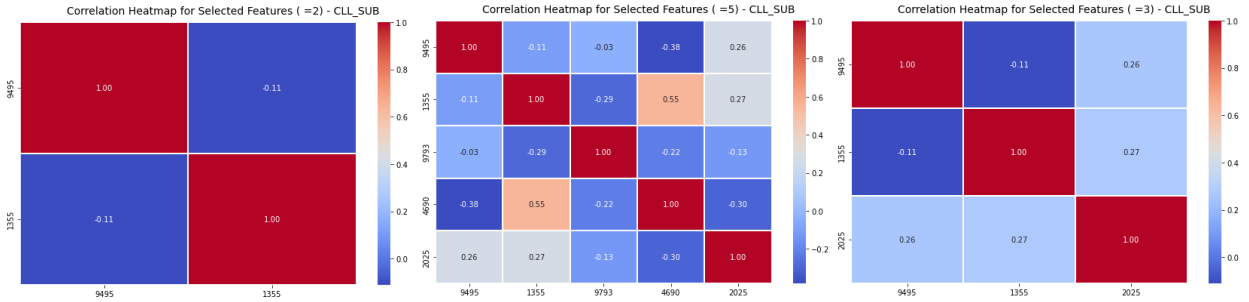


Figure-19: Correlation Heatmap for CLL_SUB

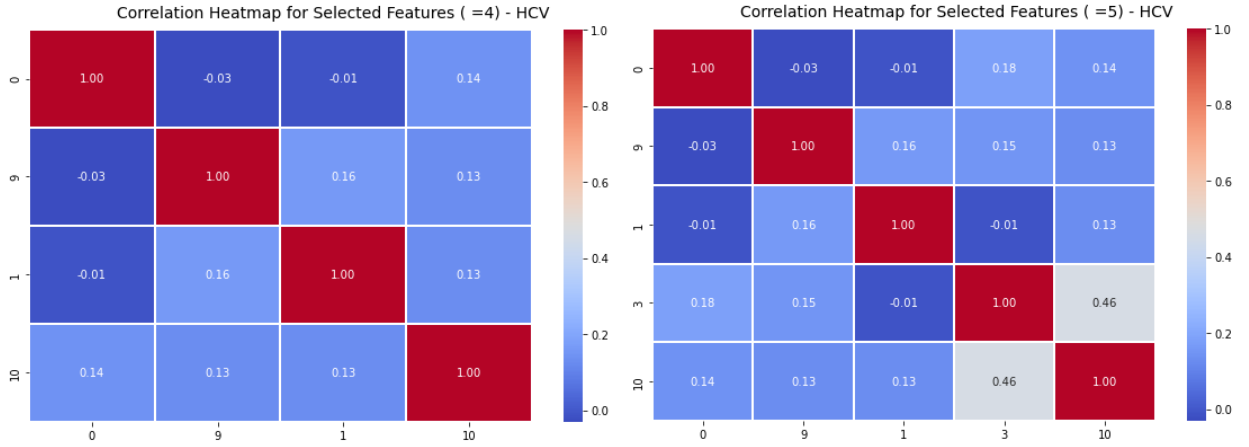


Figure-20: Correlation Heatmap for HCV

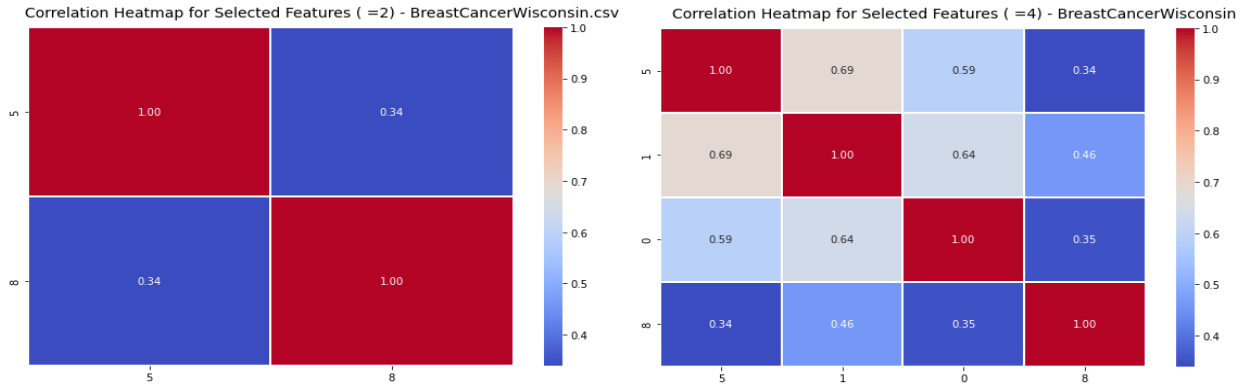


Figure-21: Correlation Heatmap for Breast Cancer Wisconsin

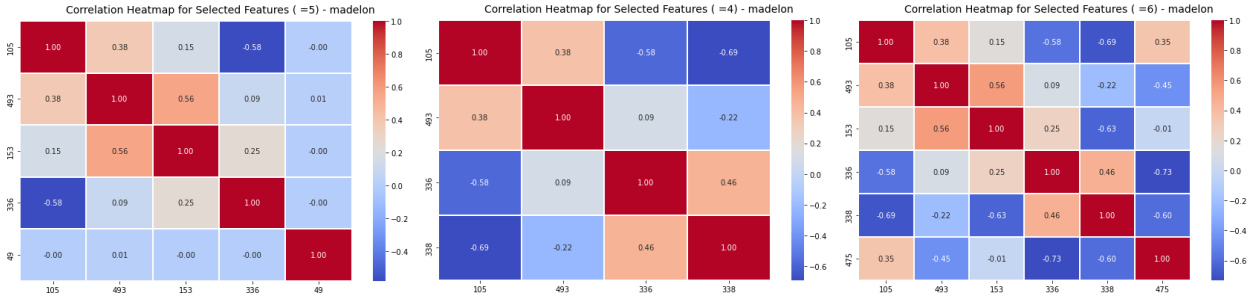


Figure-22: Correlation Heatmap for madelon

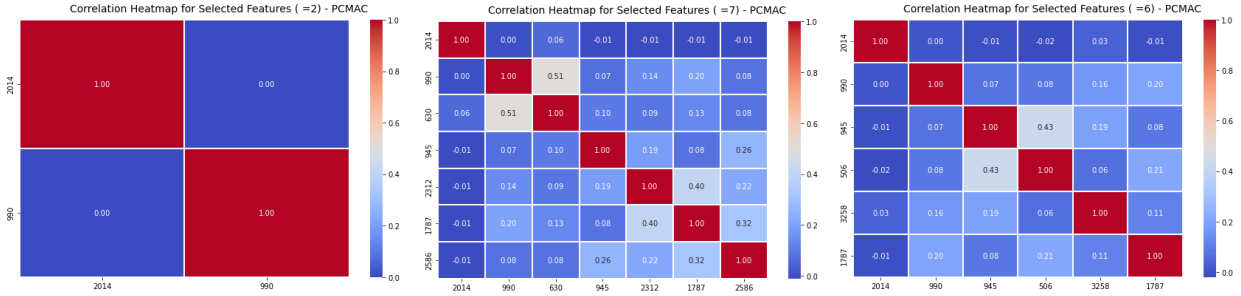


Figure-23: Correlation Heatmap for PCMAC

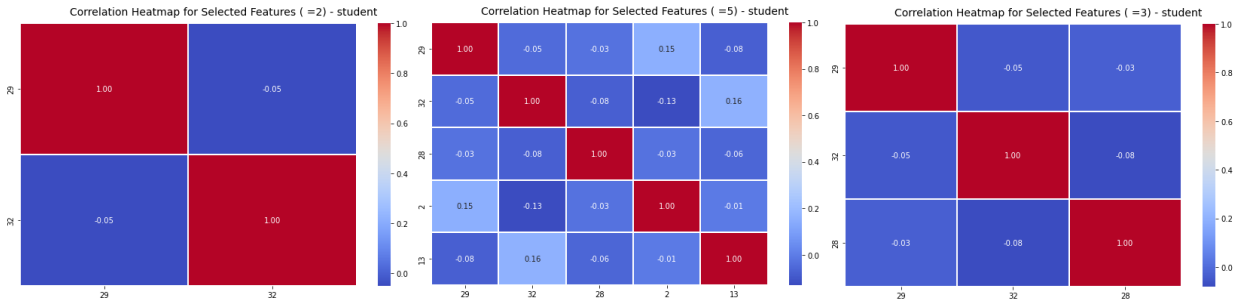


Figure-24: Correlation Heatmap for student

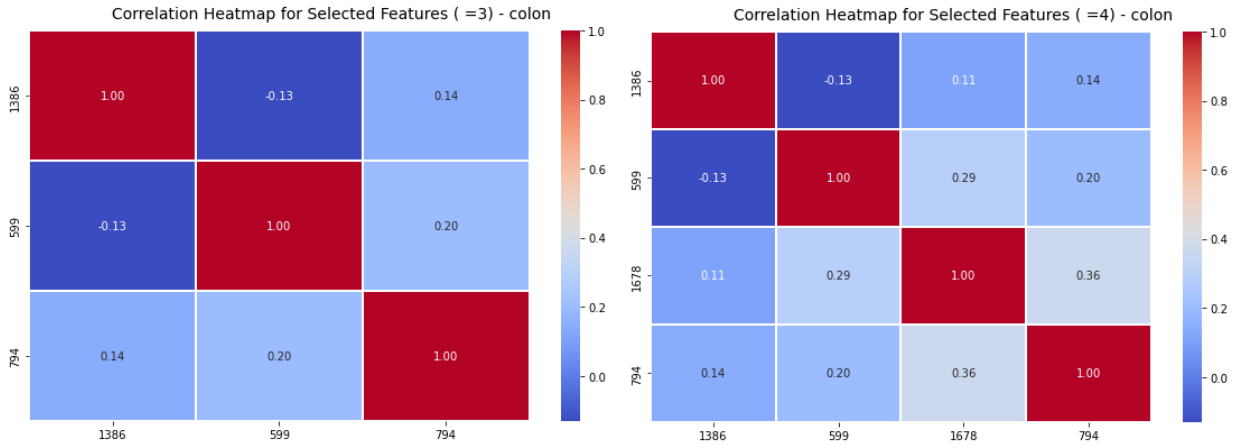


Figure-25: Correlation Heatmap for colon

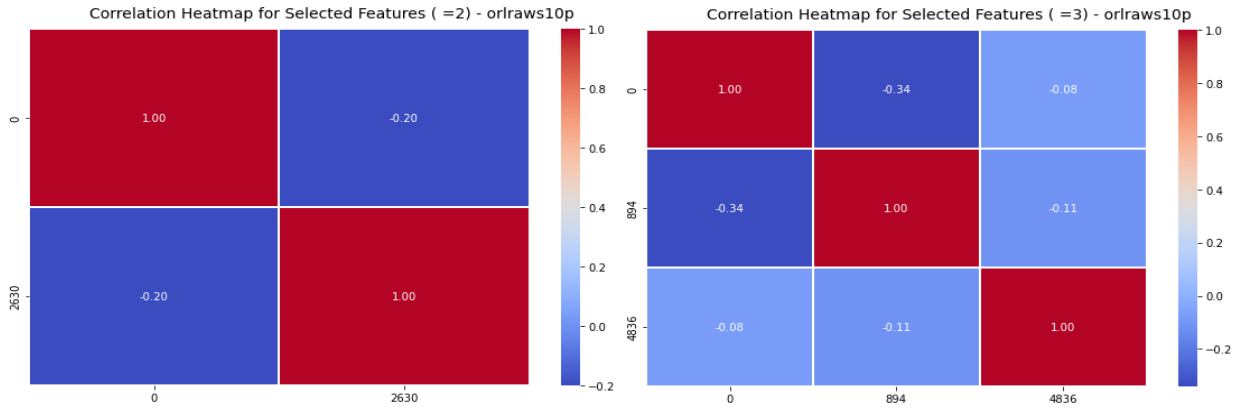


Figure-26: Correlation Heatmap for orlraws10p

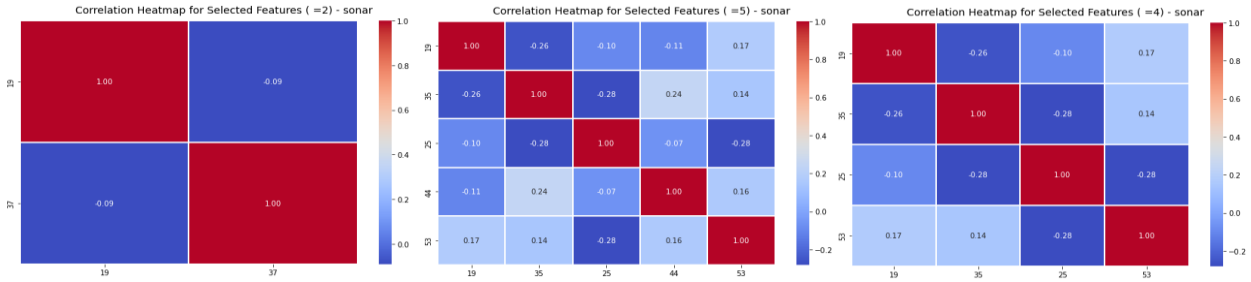


Figure-27: Correlation Heatmap for sonar

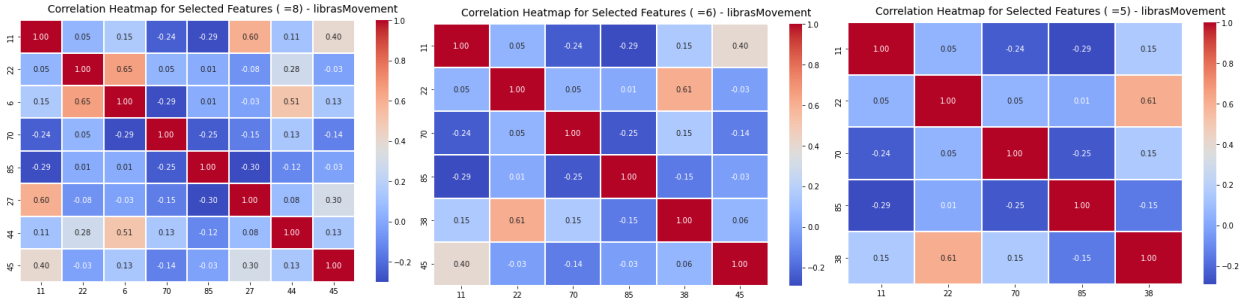


Figure-28: Correlation Heatmap for librasMovement

4.4 Selected features

Table-2: Selected Features

Data	CVI Methods on Feature Points	Learned number of Selective Features	Selective Feature Set (first column starts with index zero) arranged as per LS
diabetes	SIL	2	[0, 2]
	DIS	2	[0, 2]
	ITE	2	[0, 2]
arcene	SIL	2	[7201, 4483]
	DIS	4	[7201, 4483, 5017, 8136]
	ITE	4	[7201, 4483, 5017, 8136]
CLL_SUB	SIL	2	[9495, 1355]
	DIS	5	[9495, 1355, 1793, 4690, 2025]
	ITE	3	[9495, 1355, 2025]
HCV	SIL	4	[0, 9, 1, 10]
	DIS	5	[0, 9, 1, 3, 10]
	ITE	4	[0, 9, 1, 10]
BreastCancerWisconsin	SIL	2	[5, 8]
	DIS	4	[5, 1, 0, 8]
	ITE	4	[5, 1, 0, 8]
madelon	SIL	5	[105, 493, 153, 336, 49]
	DIS	4	[105, 493, 336, 338]

	ITE	6	[105, 493, 153, 336, 338, 475]
PCMAC	SIL	2	[2014, 990]
	DIS	7	[2014, 990, 630, 945, 2312, 1787, 2586]
	ITE	6	[2014, 990, 945, 506, 3258, 1787]
student	SIL	2	[29, 32]
	DIS	5	[29, 32, 28, 2, 13]
	ITE	3	[29, 32, 28]]
colon	SIL	3	[1386, 599, 794]
	DIS	4	[1386, 599, 1678, 794]
	ITE	4	[1386, 599, 1678, 794]
orlraws10p	SIL	2	[0, 2630]
	DIS	3	[0, 894, 4836]
	ITE	3	[0, 894, 4836]
sonar	SIL	2	[19, 37]
	DIS	5	[19, 35, 25, 44, 53]
	ITE	4	[19, 35, 25, 53]
librasMovement	SIL	8	[11, 22, 6, 70, 85, 27, 44, 45]
	DIS	6	[11, 22, 70, 85, 38, 45]
	ITE	5	[11, 22, 70, 85, 38]

Table-2 shows the learned number of selective features found in different CVI methods and the respective selective feature set which provides the original column index. Column index starts with zero i.e., first column is indexed as zero this facilitates the users of this data to ensure the selective column index to be used intact in most of the data science programming tools like python etc.

4.5 Accuracy

Two different measurement strategies are taken to check the accuracy of selected features. First (Table-3), we use initial labels (found in data source) to differentiate classification accuracy between all features and selected features. Second (Table-4), we produce new labels by clustering the data points of data set using renowned k-means technique. We have not supplied the cluster numbers of k-means arbitrarily rather we have used our Directive Elbow (depicted in 3.1) method to find the learned optimal number of clusters in data points. Then we generate the new labels in data points of optimal number of clusters as provided by the Directive Elbow. Now we use new labels to differentiate classification accuracy between all features and selected features. For classification, we have used two renowned methods KNN (K=5) and SVM. We use tenfold cross validation and average results are displayed.

TABLE-3: Accuracy test using initial labels

Data	Base Accuracy with Initial Labels (KNN)	Base Accuracy with Initial Labels (SVM)	CVI Methods on Feature Points	Learned number of Selective Features	Selective Feature's Accuracy With Initial Labels (KNN)	Selective Feature's Accuracy With Initial Labels (SVM)
diabetes	86	93	SIL	2	82	86
			DIS	2		

			ITE	2		
arcene	86	88	SIL	2	76	68
			DIS	4	75	65
			ITE	4		
CLL_SUB	49	72	SIL	2	63	51
			DIS	5	50	62
			ITE	3	59	55
HCV	91	95	SIL	4	90	89
			DIS	5	91	91
			ITE	4	90	89
BreastCancerWisconsin	97	97	SIL	2	97	97
			DIS	4	97	96
			ITE	4		
madelon	73	56	SIL	5	78	59
			DIS	4	80	61
			ITE	6	82	60
PCMAC	75	87	SIL	2	54	55
			DIS	7	65	61
			ITE	6	66	72
student	86	91	SIL	2	86	89
			DIS	5	85	90
			ITE	3	85	90
colon	74	77	SIL	3	77	80
			DIS	4	83	77
			ITE	4		
orlraws10p	89	97	SIL	2	58	36
			DIS	3	81	75
			ITE	3		
sonar	79	71	SIL	2	73	63
			DIS	5	72	78
			ITE	4	70	69
librasMovement	73	82	SIL	8	73	66
			DIS	6	65	55
			ITE	5	60	45

TABLE-4: Accuracy test using new labels

Data	CVI Methods on Data Points	Learned number of Clusters on Data Points	All Feature's Accuracy With new Labels (KNN)	Selective Feature's Accuracy With new Labels (KNN)	All Feature's Accuracy With new Labels (SVM)	Selective Feature's Accuracy With new Labels (SVM)
diabetes	SIL	2	100	100	99	99

	DIS	4	98	99	96	97
	ITE	3	100	100	98	98
arcene	SIL	3	100	100	100	100
	DIS	3				
	ITE	3				
CLL_SUB	SIL	2	100	100	100	99
	DIS	3	99	99	95	98
	ITE	3				
HCV	SIL	2	99	99	100	99
	DIS	5	96	91	97	94
	ITE	4	97	95	100	99
BreastCancerWisconsin	SIL	2	99	99	99	98
	DIS	2				
	ITE	2				
madelon	SIL	2	87	94	82	98
	DIS	5				
	ITE	5				
PCMAC	SIL	2	99	99	100	100
	DIS	9	99	99	99	99
	ITE	6	99	99	99	99
student	SIL	2	94	97	95	96
	DIS	5	90	92	90	92
	ITE	5	89	92	90	93
colon	SIL	2	87	61	93	73
	DIS	3				
	ITE	3				
orlraws10p	SIL	3	99	86	100	84
	DIS	6				
	ITE	6				
sonar	SIL	2	91	82	94	82
	DIS	4				
	ITE	4				
librasMovement	SIL	10	93	90	96	90
	DIS	6				
	ITE	6				

TABLE-5: Best Accuracies and Drops using Initial Label

Data	Number of Features	Best Accuracy with All Features	Number of Selected Features	Best Accuracy with Selected Features	Drop (%) in number of Features	Drop (%) in Accuracy (-ve indicates gain)

diabetes	16	93	2	86	87.5	7
arcene	10000	88	2	76	99.98	12
CLL_SUB	11340	72	2	63	99.98	9
HCV	12	95	5	91	58.33	4
BreastCancerWisconsin	9	97	2	97	77.78	0
madelon	500	73	6	82	98.8	-9
PCMAC	3289	87	6	72	99.82	15
student	33	91	3	90	90.91	1
colon	2000	77	4	83	99.8	-6
orlraws10p	10305	97	3	81	99.97	16
sonar	60	79	5	78	91.67	1
librasMovement	90	82	8	73	91.11	9

Figure-29

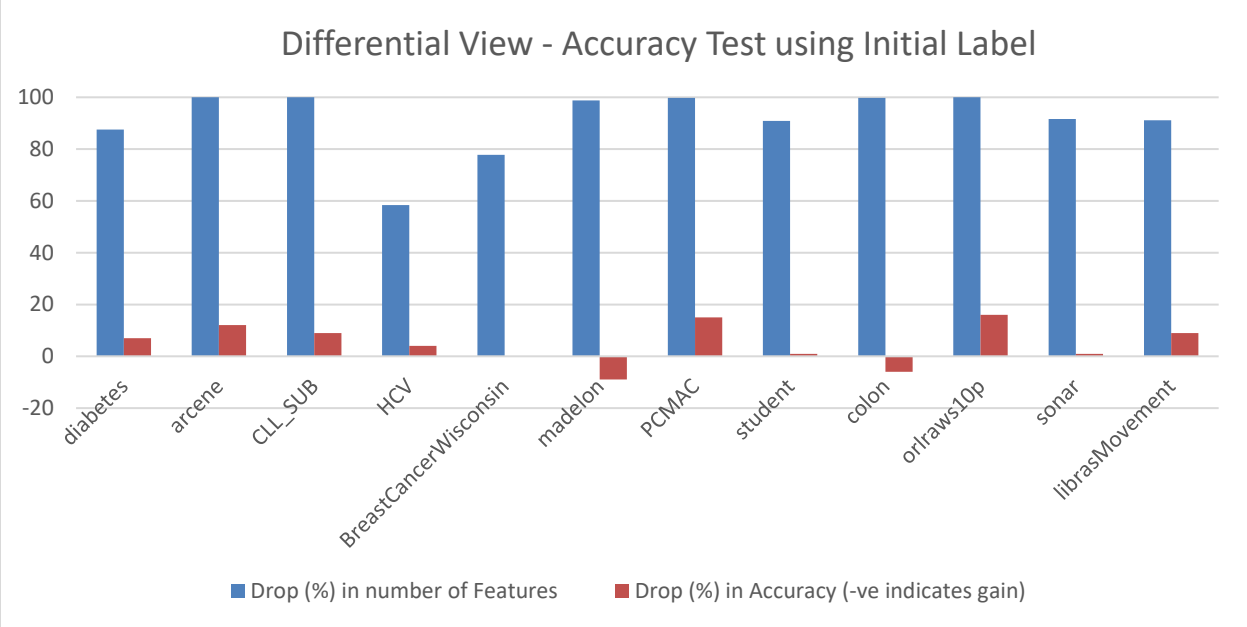
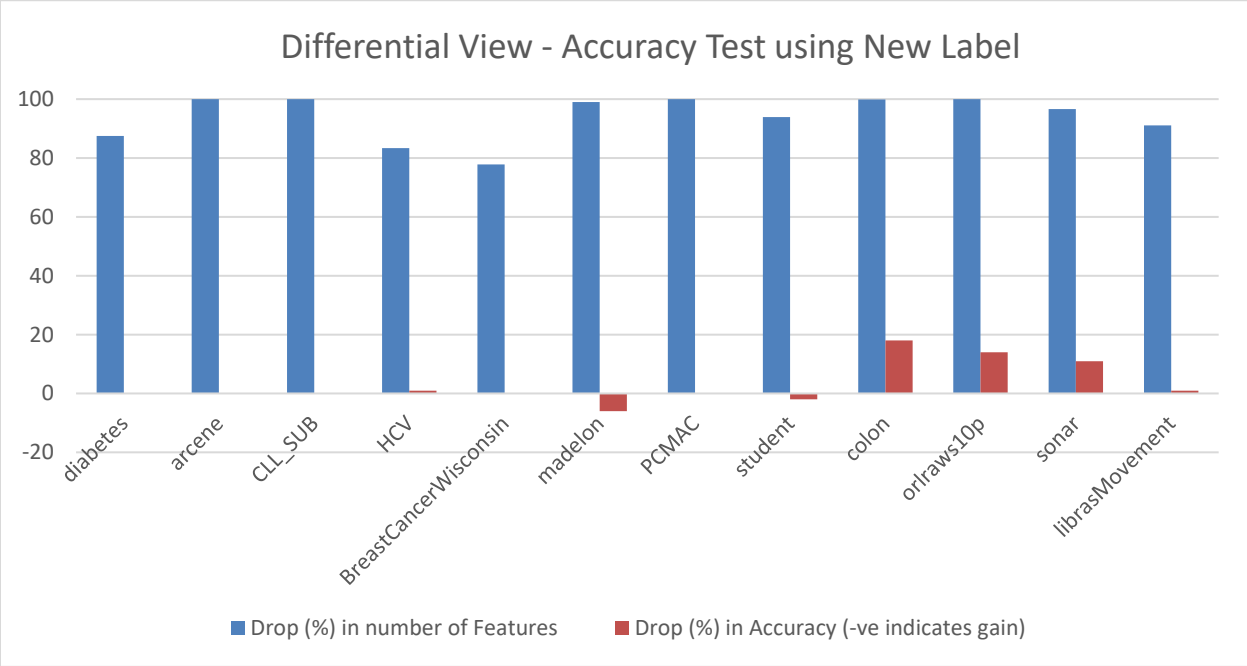


Table-5 shows the difference of best accuracies with all features and selected features found in Table-3. We call it drop percentage in accuracy. Negatives in drop percentage in accuracy means accuracy with selected features is increased with respect to the accuracy with all features. Drop percentage in number of features means the percentage of discarded features which is significantly very high as proposed in the paper. Here all the accuracies are computed using the labels found in the data source i.e., initial labels. Figure-29 represents the bar histogram of drop in numbers of features and drop in accuracy using initial label. From the figure-29 we have observed the drop in percentage of features is significantly very high for all the cases and the drop in accuracy is not significantly high (within ten percent) in almost all the cases except few cases like orlraws10p (16%), PCMAC (15%) & arcene (12%). We have achieved gain in accuracy in two cases in madelon & colon data sets.

TABLE-6: Best Accuracies and Drops using New Label

Data	Number of Features	Best Accuracy With All Features	Number of Selected Features	Best Accuracy With Selected Features	Drop (%) in number of Features	Drop (%) in Accuracy (-ve indicates gain)
diabetes	16	100	2	100	87.5	0
arcene	10000	100	2	100	99.98	0
CLL_SUB	11340	100	2	100	99.98	0
HCV	12	100	2	99	83.33	1
BreastCancerWisconsin	9	99	2	99	77.78	0
madelon	500	93	5	99	99	-6
PCMAC	3289	100	2	100	99.94	0
student	33	95	2	97	93.94	-2
colon	2000	96	3	78	99.85	18
orlraws10p	10305	100	3	86	99.97	14
sonar	60	96	2	85	96.67	11
librasMovement	90	97	8	96	91.11	1

Figure-30



In Table-6 best accuracies are taken from Table-4 and the accuracies are computed using new labels. New class labels are provided by the K-means using learned optimal number of clusters in data points. Both the accuracies, with all features and with selected features are enhanced with new labels indicates that the respective clustering on data points is sufficiently good with learned optimal number of clusters. Figure-30 indicates that the drops in accuracy are vanished in almost all cases except colon (18%), orlraws10p (14%) & sonar (11%) datasets and the drop in percentage

of features is significantly very high for all the cases. We have achieved gain in accuracy in two cases in madelon & student data sets.

5. Conclusion & Future work

This model selects non redundant and relevant features from a dataset and test accuracy of selected features is found satisfactory. This model does not require label of data or number of selected features to select redundant and relevant features. It depends on the directive elbow method to learn the optimal number of clusters in a feature set so that we cluster features efficiently and removal of redundant features is ensured. Picking up most relevant feature from a feature cluster depends on the Laplacian Score of features among the members of the cluster. Best feature from a cluster is being selected and every cluster provides one best feature. Thus, we get number of selected features automatically equals to the optimal number of clusters in a feature set.

We observed significant drops (ten to twenty percentage) in accuracy with selected feature set in few cases. This may be due to some leftover important features in a feature cluster as representative features which are not included in the selected feature set. We selected only one representative feature (the best one) from each feature clusters. In our future study of research, selection of more than one representative features from a feature cluster may be required to enhance drops in accuracy in this scenario.

This model provides numbers of experimental results like selected feature sets in twelve real life datasets with their original feature index, correlation heatmap of selected features, and elbow curves of three different coefficient (CVI) measures computed under varying K of K-means. These data may help researchers in this field.

Compliance with ethical standards

Conflict of Interest: The authors declare that there is no conflict of interest regarding the publication of this paper.

Ethical approval: This article does not contain any studies with human participants or animals performed by any of the authors.

Informed consent: Informed consent was obtained from all individual participants included in the study.

Data availability statements: Data is available from the authors upon reasonable request.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] He Xiaofei, Deng Cai, and Partha Niyogi, "Laplacian score for feature selection | Proceedings of the 18th International Conference on Neural Information Processing Systems," in *MIT Press Cambridge, MA, United States*, 2005, pp. 507–514.

[2] Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification, 2nd Edition | Wiley*.

[3] Dash M. and Liu H., "Feature selection for classification," *Intelligent Data Analysis*, vol. 1, no. 1–4, pp. 131–156, Jan. 1997, doi: 10.1016/S1088-467X(97)00008-5.

[4] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artif Intell*, vol. 97, no. 1–2, pp. 273–324, Dec. 1997, doi: 10.1016/S0004-3702(97)00043-X.

[5] J. Saha and J. Mukherjee, "CNAK: Cluster number assisted K-means," *Pattern Recognit*, vol. 110, p. 107625, Feb. 2021, doi: 10.1016/J.PATCOG.2020.107625.

[6] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J Comput Appl Math*, vol. 20, no. C, pp. 53–65, Nov. 1987, doi: 10.1016/0377-0427(87)90125-7.

[7] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Pérez, and I. Perona, "An extensive

comparative study of cluster validity indices," *Pattern Recognit*, vol. 46, no. 1, pp. 243–256, Jan. 2013, doi: 10.1016/J.PATCOG.2012.07.021.

[8] C. Subbalakshmi, G. Rama Krishna, K. M. Rao, and V. Rao, "A Method to Find Optimum Number of Clusters Based on Fuzzy Silhouette on Dynamic Data Set," *Procedia Comput Sci*, vol. 46, pp. 346–353, 2015, doi: 10.1016/j.procs.2015.02.030.

[9] Jr. David J. Ketchen and Christopher L. Shook, "The Application of Cluster Analysis in Strategic Management Research: An Analysis and Critique on JSTOR," *Strategic Management Journal Vol. 17, No. 6: Wiley*, Jun. 1996. https://www.jstor.org/stable/2486927 (accessed Mar. 24, 2022).

[10] P. Mitra, C. A. Murthy, and S. K. Pal, "Unsupervised feature selection using feature similarity," *IEEE Trans Pattern Anal Mach Intell*, vol. 24, no. 3, pp. 301–312, Mar. 2002, doi: 10.1109/34.990133.

[11] Y. M. Cheung and H. Jia, "Unsupervised feature selection with feature clustering," *Proceedings - 2012 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2012*, pp. 9–15, 2012, doi: 10.1109/WI-IAT.2012.259.

[12] X. Yan, S. Nazmi, B. A. Erol, A. Homaifar, B. Gebru, and E. Tunstel, "An efficient unsupervised feature selection procedure through feature clustering," *Pattern Recognit Lett*, vol. 131, pp. 277–284, Mar. 2020, doi: 10.1016/J.PATREC.2019.12.022.

[13] S. Bandyopadhyay, T. Bhadra, P. Mitra, and U. Maulik, "Integration of dense subgraph finding with feature clustering for unsupervised feature selection," *Pattern Recognit Lett*, vol. 40, no. 1, pp. 104–112, Apr. 2014, doi: 10.1016/J.PATREC.2013.12.008.

[14] G. Alfian *et al.*, "Predicting Breast Cancer from Risk Factors Using SVM and Extra-Trees-Based Feature Selection Method," *Computers 2022, Vol. 11, Page 136*, vol. 11, no. 9, p. 136, May 2022, doi: 10.3390/COMPUTERS11090136.

[15] M. Zivkovic, C. Stoean, A. Chhabra, N. Budimirovic, A. Petrovic, and N. Bacanin, "Novel Improved Salp Swarm Algorithm: An Application for Feature Selection," *Sensors (Basel)*, vol. 22, no. 5, May 2022, doi: 10.3390/S22051711.

[16] E. S. M. El-Kenawy *et al.*, "Novel Meta-Heuristic Algorithm for Feature Selection, Unconstrained Functions and Engineering Problems," *IEEE Access*, vol. 10, pp. 40536–40555, 2022, doi: 10.1109/ACCESS.2022.3166901.

[17] B. Abdollahzadeh and F. S. Gharehchopogh, "A multi-objective optimization algorithm for feature selection problems," *Eng Comput*, vol. 38, no. 3, pp. 1845–1863, May 2022, doi: 10.1007/S00366-021-01369-9/TABLES/7.

[18] L. Hu, L. Gao, Y. Li, P. Zhang, and W. Gao, "Feature-specific mutual information variation for multi-label feature selection," *Inf Sci (N Y)*, vol. 593, pp. 449–471, May 2022, doi: 10.1016/J.INS.2022.02.024.

[19] A. Tiwari and A. Chaturvedi, "A hybrid feature selection approach based on information theory and dynamic butterfly optimization algorithm for data classification," *Expert Syst Appl*, vol. 196, p. 116621, May 2022, doi: 10.1016/J.ESWA.2022.116621.

[20] M. H. Nadimi-Shahraki, H. Zamani, and S. Mirjalili, "Enhanced whale optimization algorithm for medical feature selection: A COVID-19 case study," *Comput Biol Med*, vol. 148, p. 105858, May 2022, doi: 10.1016/J.COMPBIOMED.2022.105858.

[21] P. Zhu, X. Hou, K. Tang, Y. Liu, Y. P. Zhao, and Z. Wang, "Unsupervised feature selection through combining graph learning and ℓ2,0-norm constraint," *Inf Sci (N Y)*, vol. 622, pp. 68–82, May 2023, doi: 10.1016/J.INS.2022.11.156.

[22] A. Alsahaf, N. Petkov, V. Shenoy, and G. Azzopardi, "A framework for feature selection through boosting," *Expert Syst Appl*, vol. 187, 2022, doi: 10.1016/j.eswa.2021.115895.

[23] A. Wahid, D. M. Khan, I. Hussain, S. A. Khan, and Z. Khan, "Unsupervised feature selection with robust data reconstruction (UFS-RDR) and outlier detection," *Expert Syst Appl*, vol. 201, 2022, doi: 10.1016/j.eswa.2022.117008.

[24] P. Huang and X. Yang, "Unsupervised feature selection via adaptive graph and dependency score," *Pattern Recognit*, vol. 127, 2022, doi: 10.1016/j.patcog.2022.108622.

[25] S. Ouadfel and M. A. Elaziz, "Efficient high-dimension feature selection based on enhanced equilibrium optimizer," *Expert Syst Appl*, vol. 187, 2022, doi: 10.1016/j.eswa.2021.115882.

[26] D. Patel, A. K. Saxena, S. Laha, and G. M. Ansari, "A Novel Scheme for Feature Selection Using Filter Approach," *Proceedings of the 2022 7th International Conference on Computing, Communication and Security, ICCCS 2022 and 2022 4th International Conference on Big Data and Computational Intelligence, ICBDCI 2022*, 2022, doi: 10.1109/ICCCS55188.2022.10079604.

[27] J. S. Wu, J. X. Liu, J. Y. Wu, and W. Huang, "Dictionary learning for unsupervised feature selection via dual sparse regression," *Applied Intelligence*, pp. 1–17, Feb. 2023, doi: 10.1007/S10489-023-04480-0/FIGURES/8.

[28] A. Isaac, H. K. Nehemiah, S. D. Dunston, and A. Kannan, "Feature selection and classification using bio-inspired algorithms for the diagnosis of pulmonary emphysema subtypes," *Int J Imaging Syst Technol*, 2023, doi: 10.1002/IMA.22867.

[29] A. A. Abdulhussien, M. F. Nasrudin, S. M. Darwish, and Z. Abdi Alkareem Alyasseri, "Feature selection method based on quantum inspired genetic algorithm for Arabic signature verification," *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 3, pp. 141–156, Mar. 2023, doi: 10.1016/J.JKSUCI.2023.02.005.

[30] S. Agrawal, A. Tiwari, B. Yaduvanshi, and P. Rajak, "Feature subset selection using multimodal multiobjective differential evolution," *Knowl Based Syst*, vol. 265, p. 110361, Apr. 2023, doi: 10.1016/J.KNOSYS.2023.110361.

[31] M. Zivkovic, C. Stoean, A. Chhabra, N. Budimirovic, A. Petrovic, and N. Bacanin, "Novel Improved Salp Swarm Algorithm: An Application for Feature Selection," *Sensors (Basel)*, vol. 22, no. 5, Mar. 2022, doi: 10.3390/S22051711.

[32] A. Saxena *et al.*, "A Novel Unsupervised Feature Selection Approach Using Genetic Algorithm on Partitioned Data," *Advances in Artificial Intelligence and Machine Learning*, vol. 02, no. 04, pp. 500–515, 2023, doi: 10.54364/AAIML.2022.1134.

[33] D. Chugh, H. Mittal, A. Saxena, R. Chauhan, E. Yafi, and M. Prasad, "Augmentation of Densest Subgraph Finding Unsupervised Feature Selection Using Shared Nearest Neighbor Clustering," *Algorithms 2023, Vol. 16, Page 28*, vol. 16, no. 1, p. 28, Jan. 2023, doi: 10.3390/A16010028.

[34] D. Patel, A. Kumar Saxena, S. Laha, R. Prasad, and U. Roy, "An Exhaustive Wrapper Method for Feature Selection in

Large Dimensional Datasets (WFS)," *The Ciência & Engenharia - Science & Engineering Journal*, vol. 11, no. 1, pp. 2206–2225, Feb. 2023, doi: 10.52783/CIENCENG.V11I1.397.

[35] F. R. Chung, "Spectral graph theory. Vol. 92. American Mathematical Soc.," 1997.

[36] Dhabliya, D., Dubey, M. K., Gupta, A., & Reddy, D. H. (2022). A review on comparison of machine learning algorithms for text classification. Paper presented at the Proceedings of 5th International Conference on Contemporary Computing and Informatics, IC3I 2022, 1818-1823. doi:10.1109/IC3I56241.2022.10072502 Retrieved from www.scopus.com

[37] "UCI Machine Learning Repository." https://archive.ics.uci.edu/ml/index.php.

[38] Dhanikonda, S. R., Sowjanya, P., Ramanaiah, M. L., Joshi, R., Krishna Mohan, B. H., Dhabliya, D., & Raja, N. K. (2022). An efficient deep learning model with interrelated tagging prototype with segmentation for telugu optical character recognition. Scientific Programming, 2022 doi:10.1155/2022/1059004 Dhingra, M.,

[39] "Datasets | Feature Selection @ ASU." https://jundongl.github.io/scikit-feature/OLD/datasets_old.html.