Artificial Intelligence Review (2021) 54:3473–3515
https://doi.org/10.1007/s10462-020-09928-0

Various dimension reduction techniques for high dimensional data analysis: a review

Papia Ray[1] · S. Surender Reddy[2] · Tuhina Banerjee[1]

Published online: 8 January 2021

Abstract
In the era of healthcare, and its related research fields, the dimensionality problem of high dimensional data is a massive challenge as it contains a huge number of variables forming complex data matrices. The demand for dimension reduction of complex data is growing immensely to improvise data prediction, analysis and visualization. In general, dimension reduction techniques are defined as a compression of dataset from higher dimensional matrix to lower dimensional matrix. Several computational techniques have been implemented for data dimension reduction, which is further segregated into two categories such as feature extraction and feature selection. In this review, a detailed investigation of various feature extraction and feature selection methods has been carried out with a systematic comparison of several dimension reduction techniques for the analysis of high dimensional data and to overcome the problem of data loss. Then, some case studies are also cited to verify the better approach for data dimension reduction by considering few advances described in the technical literature. This review paper may guide researchers to choose the most effective method for satisfactory analysis of high dimensional data.

Keywords Canonical correlation analysis · Feature extraction · Feature selection · Local Fisher's discriminate analysis · Locally linear embedding · Principle component analysis

✉ S. Surender Reddy
surender@wsu.ac.kr

Papia Ray
papia_ray@yahoo.co.in

Tuhina Banerjee
tuhinabanerjee97@gmail.com

[1] Department of Electrical Engineering, Veer Surendra Sai University of Technology, Burla, Sambalpur, India

[2] Department of Rail Road and Electrical Engineering, Woosong University, Daejeon, Republic of Korea

1 Introduction

In the era of rapid technology advancement, it is important to handle the data properly. Otherwise the chance of data loss occurs. Sometimes the extracted features are high in dimensionality and tough to handle. It is also problematic to classify such high dimension data. Because of high dimensional datasets, some redundant features occurs which leads to data loss. Such kind of high dimension data (HDD) cannot be handled through regular database management. In this paper, different feature extraction and feature selection techniques have been investigated to reduce the data appropriately for further classification. The significance of HDD analysis is improving from time to time. The HDD is subset of big data that is a collection of huge amount of data which can be analysed to find useful information and patterns.

Main purpose of this survey is to guide the researchers to choose the most effective method for satisfactory analysis of high dimensional data. Data used in classification should be prominent, otherwise it will not show the exact result or it can further lead to data loss. So for proper data handling and to make sure that data will not be lost during the experiment, appropriate reduction techniques have to be considered. A brief discussion about the reduction techniques for health care big data analysis is carried out in this paper citing some cases.

Recently a term called "high dimensional data (HDD)" is a buzzword in medical science, data science and healthcare sectors (Alexander and Wang 2017; Hossain and Muhammad 2016). Its application has tremendous impact on data analysis, visualization, processing and classification. Huge amount of patient data can be recorded which could be utilized by machine learning for the benefit of health care sector (Archenaa and Mary Anita 2015). A dataset represents a statistical data matrix with domains or subjects in rows and variables in columns. An individual column from the input dataset is termed as features. Technically a feature is the measurable properties of sampled data. The HDD depends on three factors such as data velocity (rate at which data is generated), data veracity (types of data) and data volume. It is useful in data interpretation, management, analysis and visualization (Raghupati and Raghupati 2014). But the storage, processing and maintenance of such a mass feature of HDD needs a lot of memory space which may result in data loss. Further data privacy, global data transparency, data storage and security are some of the unavoidable issues in research fields of data analysis (Deyan and Zhao 2012). This problem leads to an inescapable issue known as "Curse of Dimensionality" (Van der Linden and Dufresne 2017) and may increase the volume of data vulnerably that tends to form a sparse data. When HDD is analyzed either mathematically or statistically, the sparsity creates problem. A sparse dataset contains a large variable matrix; however, the matrix's cell may not contain actual data. These cells are filled up with numeric '0's only, due to which it may occupy more memory storage resulting in wastage of memory. This may be the key factor for data loss (Al-Bakri and Soukaena 2018). Hence in those cases, it is extremely necessary to reduce the dimension or attributes of data by dimension reduction techniques (DRT). In general, DRT maps the higher order dimensions into two-dimensional or three-dimensional form without affecting the salient features for the analysis of data (Tan 2018). At first the large dimension of data is managed to reduce it into a lower dimensional space retaining the originality of data attributes. In the next step reduced form of data is sent for processing and analysis. Figure 1 summarizes the whole concept of DRT in a block diagram (Sacha et al. 2017).

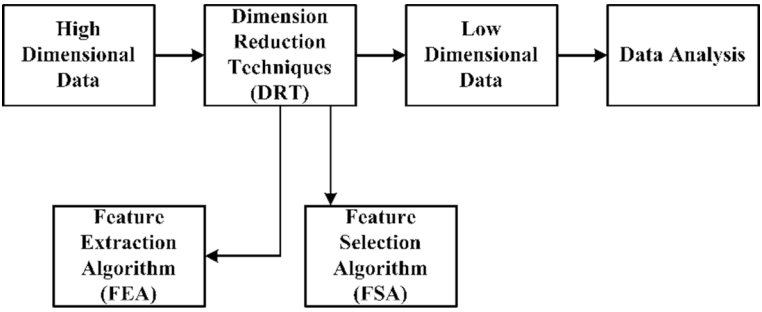

Fig. 1 Block diagram showing DRT concept

From Fig. 1, it can be observed that DRT is focused at converting the data from higher dimensional space into a lower dimensional space for better data analysis retaining most of its essential attributes in original form. To evaluate the discussion statistically, DRTs are further classified into two categories such as FEA and FSA. Since our objective is to make a complete comprehension of data, the earlier research fails to discuss some important taxonomical issues that create information gap between experts and beginners, especially the people who are just one step ahead in biomedical science and its applications. A debate is going on since many years among the researchers who have advocated that the optimal selection of feature subset leads to an accurate model; whereas others argued on finding the best optimal feature subset that include feature transformation (Jiang and Li 2015; Ozdenizci and Erdogmus 2019; Xu et al. 2017). Dimension reduction is the main topic related to this problem that refers to the transformation of high-dimensional data to a low-dimensional representation. Feature extraction is the process of transforming the raw data into mass feature subset either in time or frequency domain or both time–frequency domains. However feature selection is a prior technique of choosing some of the best features from the feature subset that boost the research area in data interpretation and analysis. Further feature selection scheme is needed when we want to determine the "best" feature subset for most approachable data anticipation (Ding et al. 2012; Chandrashekar and Sahin 2014; Kira and Rendell 1992; Liu and Motoda 2007).

From the literature survey, it can be observed that the prior objective of searching a desired feature subset is to find a group of features that can be used in analysis of the original higher dimensional dataset with effective classification accuracy. To achieve this primary aim, the sub-objective that has to be successfully fulfilled is the complex high dimensional data which must be converted into a lower dimensional data matrix. Different algorithms may call for different objectives, which may divert the primary objective of our search. Hence, our approaches in describing different algorithms should be brief and limited within some popularly used methods (Mazomenos et al. 2013; Aggarwal and Cheng 2012; McDonnell et al. 2010).

The rest sections of the paper are organized as follows. Section 2 highlights the introduction of some of the feature extraction algorithms. In Sect. 3, some of the proposed feature selection algorithm is described. Section 4 includes a comparison study and reports the overall comparative analysis. An experimental setup has been performed in Sect. 5 to validate the analysis. Eventually Sect. 6 describes the conclusion at a glance.

2 Feature extraction algorithms (FEA)

In this section, the algorithms of feature extraction have been discussed. The most former attribute for dimensional reduction technique is feature extraction (Gedik 2016). It is the most significantly used dimension reduction technique for the transformation of features. In earlier times, various traditional algorithms have been employed for extracting the required features from raw dataset for data preprocessing as shown in Fig. 2.

In Fig. 2, 'PCA' stands for principle component analysis, 'LDA' means local discriminate analysis 'CCA' is canonical correlation analysis and 'NMF' stands for non-negative matrix factorization. The preprocessing starts with data transformation, which includes four aspects such as: standardization, normalization, noise filtration and feature extraction. Let us take an example to explain the scheme easily. A data matrix 'A' with n dimension is represented as:

$$\mathrm{A} = \left\{\mathrm{a}_1, \mathrm{a}_2, \ldots, \mathrm{a}_\mathrm{n}\right\} \tag{1}$$

where $\mathrm{a}_1, \mathrm{a}_2, \ldots, \mathrm{a}_n$ are the data elements. The components 'f_i' are the features for the data vectors. The objective of data transformation is to transform the original feature vector a_i into a new feature vector $\mathrm{f}_i^{\mathrm{T}}$ with n^{T} dimension. The features may vary with a large scale although we deal with a single attribute. For an instance, let us take two features from vector as $\mathrm{f}_i = [\mathrm{f}_1, \mathrm{f}_2]$, however the scaling factors of measurement for both the vectors may vary. Hence comparing or adding those factors can't give any desirable results. This problem can be solved by scaling the features in a particular range as mentioned in Eq. (2).

$$\mathrm{f}_i^{\mathrm{T}} = \frac{(\mathrm{a}_i - \mathrm{n}_i)}{\Gamma_i} \tag{2}$$

Here Γ_i and μ_i are the standard deviation and mean for the original features 'f_i' respectively. After scaling, the feature subset can be reformed in a standardized form. This can be achieved by normalizing the data matrix 'A' by dividing with total counts to maintain accuracy and efficiency of the database. However handling large dimensional data is inefficient with clustering techniques. Hence, FEA is applied which is a dimensionality reduction

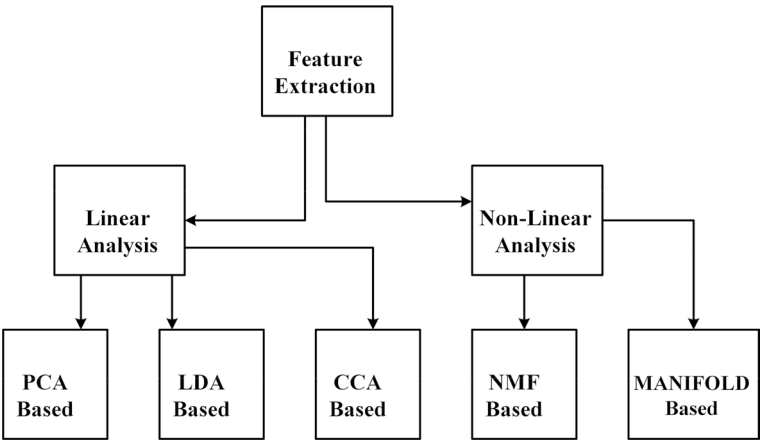

Fig. 2 Hierarchy of feature extraction method

technique that reduces the data into features and makes it a simplified one (McDonnell et al. 2010; Gedik 2016). Figure 2 depicts the hierarchy of supervised feature extraction methods. Feature extraction methods are classified in terms of components and projection. Further the components are classified in terms of linear analysis such as Principle Component Analysis (PCA), Local Fisher's Discriminate Analysis (LFDA) and Canonical Correlation Analysis (CCA). The non-linear data's are analysed via several methods such as Non-negative Matrix Factorization (NMF) and manifold learning based algorithm (Wang and Zhang 2013; Chen and Zhang 2009).

2.1 Principal component analysis (PCA)

An earlier and most popular algorithm for data dimension reduction is PCA, first developed in Behbahani et al. (2017), Kapsoulis et al. (2018). It is an orthogonal feature transformation technique that converts the correlated sampled variables into linear uncorrelated sampled variables. These new variables resembles the original variables and hence known as principal components (PC). The components are linearly combined to form various features and helps in computing maximum variance (Kapsoulis et al. 2018). The resultant is predictable and easy to understand by using the original features (Kapsoulis et al. 2018). Figure 3 shows separation of two principle components from a set of variables and indicates PCA of a dataset symbolizedwithin an elliptical subspace.

Here, axis 1 and 2 are representing 1st and 2nd principle components respectively. The extracted feature subsets are enclosed by the subspace, retaining other features outside. As mentioned in Eq. (1), observations are taken for 'f$_i$' features. Hence, the PC can be calculated as:

$$\mathrm{PC} = \sum_{i=1}^{n} x_{1i} a_i \tag{3}$$

where a$_i$ is the ith component of 'A', x_{1i} is the coefficients of PC for $i = 1, \cdots, n$. In a simple way, x_1 is the column vector, which is normalized by $x_1^{\mathrm{T}} x_1 = 1$. The variance of PC is given as the formula $x_1^{\mathrm{T}} \mathrm{S} x$, where 'S' is the covariance of the matrix. Further by maximizing the variance of PC, column vector can be determined and first PC can be denoted. The same process will be followed to find out the successive PCs (Kapsoulis et al. 2018). The algorithm of PCA is shown in Fig. 4.

Figure 4, explains the algorithm of PCA step by step. First the data matrix has been collected. In the second step, data dimension reduction process has been introduced after that the third step comprises of output of the principal components. The next step comprises of the principal components input on-line system and final step is output result.

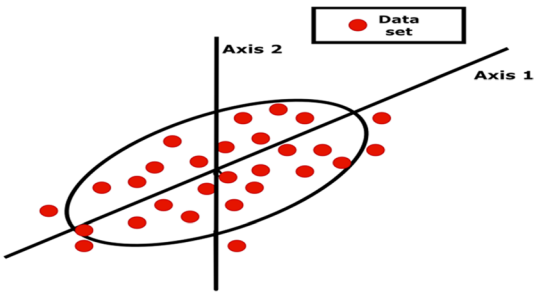

Fig. 3 Analysis of principle components

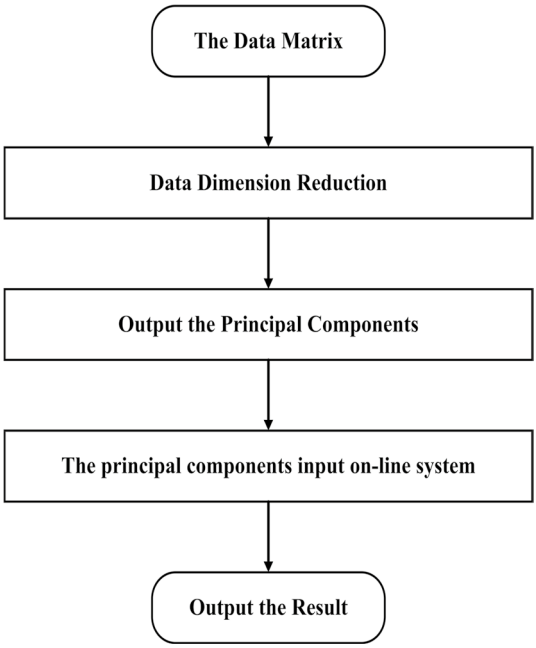

Fig. 4 Flowchart of PCA

2.2 Local fisher's discriminate analysis (LFDA)

Fisher's Discriminate Analysis (FDA) is a traditional technique of projection of data matrix with maximum discrimination, which is effective for classification performance and dimension reduction. The objective of this algorithm is to reduce the data dimensionality preserving the maximum discriminative information of the class (Zhao and Gao 2015). However, existing method gives unsatisfactory results for a multimodal class separation.

In order to reduce the dimensions of HDD, preservation of local structure of data is necessary. Hence, the concept of FDA with locality preserving projection method is merged and LFDA is formed. It separates the multimodal class structure and helps in preserving the local structure of data at the same time (Cong and Duan 2016; Li et al. 2015). The algorithm of LFDA is shown in the Fig. 5.

Figure 5 depicts the algorithm of LFDA. It shows that in first step data has been prepared than one super component has been extracted then the left information calculated. After that rank has been checked and original data set has been replaced by related statistics. After that one final component and output of discriminate component has been extracted.

2.3 Canonical correlation analysis (CCA)

The linear relationship between two variables from a multidimensional dataset is measured by a correlation-based technique, known as CCA. For each variable, it finds two bases that are optimal with respect to correlations by using multivariate analysis (Jendoubi and Strimmer 2019). Let us consider two different multivariate dataset such as

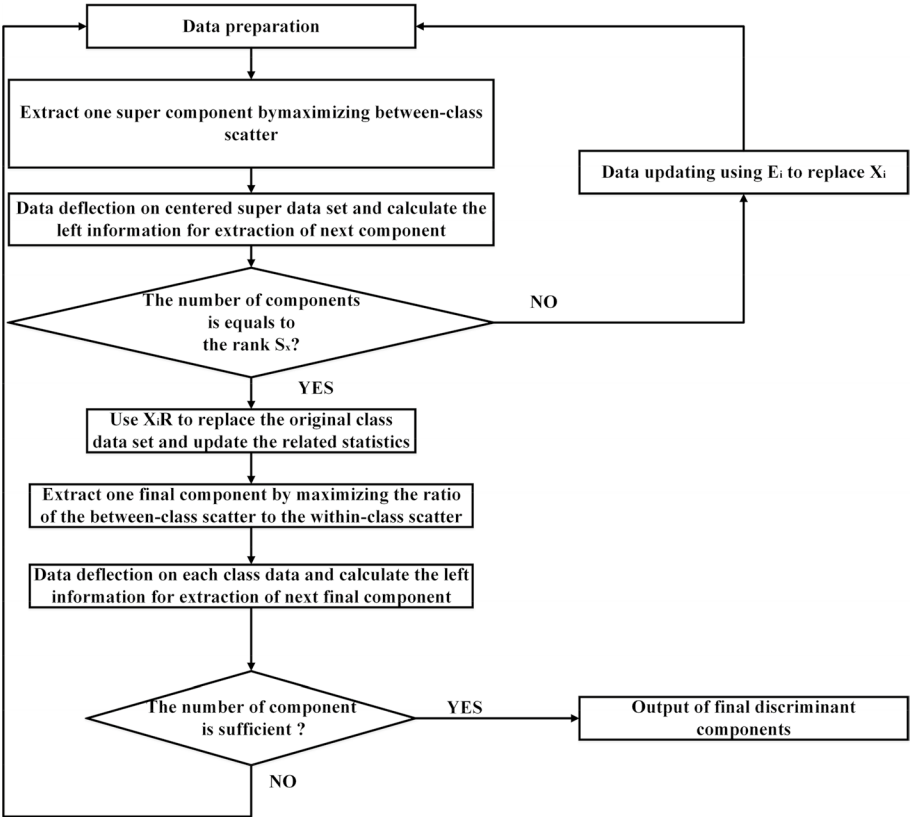

Fig. 5 Flowchart of LFDA

$$\mathrm{A} = \left\{\mathrm{a}_1, \mathrm{a}_2, \ldots, \mathrm{a}_n\right\} \text{ and } \mathrm{B} = \left\{\mathrm{b}_1, \mathrm{b}_2, \ldots, \mathrm{b}_\mathrm{n}\right\} \tag{4}$$

where $i = 1, 2, \ldots, n$ is the feature variables realization in each dataset. μ_A and μ_B be the mean vectors of 'A'and 'B', formulated as:

$$\mu_\mathrm{A} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{A} \text{ and } \mu_\mathrm{B} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{B} \tag{5}$$

The centre points can be written as $\tilde{\mathrm{A}} = \mathrm{A} - \mu_\mathrm{A}$ and $\tilde{\mathrm{B}} = \mathrm{B} - \mu_\mathrm{B}$ for datasets 'A' and 'B' respectively. After that the linear transformation of original features (variables) of the datasets into a new feature subsets such as W_A and W_B is done. The main objective of CCA is to maximize the objective function that means maximize the correlation between original feature set of 'A'and 'B'with the transformed feature set $(\mathrm{W}_\mathrm{A})^T\mathrm{A}$ and $(\mathrm{W}_\mathrm{B})^T\mathrm{B}$. The objective function is mathematically expressed as:

$$F = \frac{\mathrm{cov}((W_A)^T A,(W_B)^T B)}{\sqrt{\mathrm{var}(W_A)^T A\,\mathrm{var}(W_B)^T B}}$$
$$ÞF = \frac{(W_A)^T C_{AB}(W_B)^T}{\sqrt{((W_A)^T C_{AA} W_A)((W_B)^T C_{BB} W_B)}} \tag{6}$$

where the abbreviated terms such as 'cov' and 'var' denotes co-variance and variance between feature sets of 'A' and 'B' respectively. The term 'cov' can be expressed as $C_{AB} = \left\langle \tilde{A}\tilde{B}^T \right\rangle$ and 'var' can be expressed as similar to 'cov' as $C_{AA} = \left\langle \tilde{A}\tilde{A}^T \right\rangle$ and $C_{BB} = \left\langle \tilde{B}\tilde{B}^T \right\rangle$. The symbol $\langle . \rangle$ is referred as statistical expectation. Hence CCA for the multivariate datasets is written as,

$$\mathrm{CCA}_{AB} = \arg\max_{W_A} \max_{W_B} W_A^T C_{AB} W_B \tag{7}$$

Here, two constraints of $(W_A)^T C_{AA} W_A = 1$ and $(W_B)^T C_{BB} W_B = 1$ are taken and hence, Lagrangian functions can be determined as:

$$L_f = W_A^T C_{AB} W_B + \lambda_A \left(1 - W_A^T C_{AA} W_A\right) + \lambda_B \left(1 - W_B^T C_{BB} W_B\right) \tag{8}$$

To get the eigen values, two conditions must be satisfied such as $\frac{\delta L_f}{\delta W_A} = 0$ and $\frac{\delta L_f}{\delta W_B} = 0$.

Putting Lagrangian function in the conditions stated above, Eq. (7) can be resolved as:

$$\begin{pmatrix} 0 & C_{AB} \\ C_{BA} & 0 \end{pmatrix} \begin{pmatrix} W_A \\ W_B \end{pmatrix} = \lambda \begin{pmatrix} C_{AA} & 0 \\ 0 & C_{BB} \end{pmatrix} \begin{pmatrix} W_A \\ W_B \end{pmatrix} \tag{9}$$

where $\lambda = 2\lambda_A = 2\lambda_B$ are the eigen values (Cong and Duan 2016; Li et al. 2015; Jendoubi and Strimmer 2019). Hence, CCA involves two purposes: One is data reduction where co-variation between two sets of variables using small number of linear combinations is explained. The other is data interpretation for finding features that explains co-variation between sets of variable (Michaeli et al. 2016; Luo et al. 2015; Wilms and Croux 2015). The algorithm of CCA is shown in the Fig. 6.

Figure 6, shows that the predictors have been passed from synthetic predictor and then that result is passed through the synthetic criteria and for different other criterion. Then the result has been calculated. This algorithm has been designed to maximize the simple correlation.

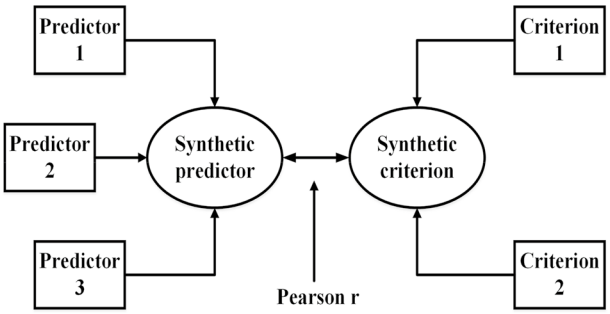

Fig. 6 Flowchart of CCA

2.4 Non-negative matrix factorization (NMF)

An innovative approach for decomposition of dataset with two non-negative factors is known as NMF based feature extraction method. This method is restricted to be non-negative for two reasons. The first reasonis based on neurophysiology in which the visual perception neurons have a non-negative firing rate (Varghese et al. 2018). The second reason is based on image processing field in which the intensity images occupy non-negative values. Let us consider a matrix A, each element a_{ij} is decomposed into a basic coefficient and decomposed coefficient of X and Y respectively. Since size of X is $i \times k$ and size of Y is $k \times j$; hence

$$a_{ij} \approx \sum X_{ik} Y_{kj} \tag{10}$$

NMF for this data matrix can be formulated by

$$\begin{cases} \min\limits_{X,Y} A = XY, X \geq 0, \quad Y \geq 0 \end{cases} \tag{11}$$

Cost function determines the level of decomposition. Here, two are selected to find out the distance between 'A' and 'XY' with a generalized KL divergence (Michaeli et al. 2016; Luo et al. 2015). The expression for KL based cost function is:

$$C_{NMF}(A||XY) \triangleq \sum_{ij} a_{ij} \ln \frac{a_{ij}}{\sum\limits_{k} X_{ik} Y_{kj}} + \sum_{k} X_{ik} Y_{kj} - a_{ij} \tag{12}$$

To minimize this expression, multiplicative update rules are applied when $X \geq 0, Y \geq 0$. The divergence is minimized to apply an algorithm to the unknown coefficients X and Y. For this matrix W is considered such that $W = XY$. Therefore, NMF based two cost functions can be expressed as:

$$\begin{cases} \min\limits_{X,Y} ||A - W||_F^2 \end{cases} \tag{13}$$

$$\begin{cases} \min\limits_{X,Y} C(A||W) \end{cases} \tag{14}$$

For $W = XY$ and $X \geq 0, Y \geq 0$.
where

$$C(A||W) = \sum_{ij} \left(A_{ij} \log \frac{A_{ij}}{W_{ij}} + A_{ij} - W_{ij} \right) \tag{15}$$

Normally the size of k is much less than the size of cost function, hence the basic coefficient is chosen to be the suitable lower-dimensional space vector. However, the outliers are treated with a robust NMF based non-square equation in $l_{2,1}$ norm as follows.

$$|A - XY||_{2,1} = \sum_{i=1}^{N} \sqrt{\sum_{j=1}^{C} (A - XY)_{j,i}^2} \tag{16}$$

Equation (16) ensures that no outliers will dominate the objective function of NMF based algorithm. Hence, a standardized NMF algorithm can be used as a novel feature extraction scheme in various applications like pattern recognition, text classification, tumor classification etc. But it has several disadvantages such as optimization problem, subspace selection, nonlinear non-negative features etc. (Wang et al. 2018a, b; Li and Ngom 2013; Zeynep et al. 2011). The algorithm of NMF is shown in the Fig. 7.

Figure 7, shows in the first step the relevant symptom has been removed. In second step the redundant symptom based on NMF has been identified. In the third step the redundant symptoms by group has been transformed. In the last step the feature subset has been selected.

2.5 Manifold learning based algorithm

Manifold learning based feature extraction method is applicable for nonlinear data dimension reduction. Nonlinear methods provide mapping of data from high dimensional attribute to low dimensional attribute or vice versa. They also visualize the proximity of data based on the distance measurement. Hence, manifold based algorithm predicts the manifolds of lower dimensional space from higher dimensional point of view and enforces to uncover those manifolds (Zhang et al. 2010; Cannistraci et al. 2010). This algorithm may acquire an internal model of the data and can be employed to map those points that are unavailable during training time. Some of the manifold based algorithms for dimension reduction are Laplacian Eigen map (LE), Locally Linear Embedding (LLE) and Isomap.

2.5.1 Isomap

When a high-dimensional data is located on or near the curved manifolds, the Euclidian distance between the vertices avoids the neighborhood distribution. To get rid of this

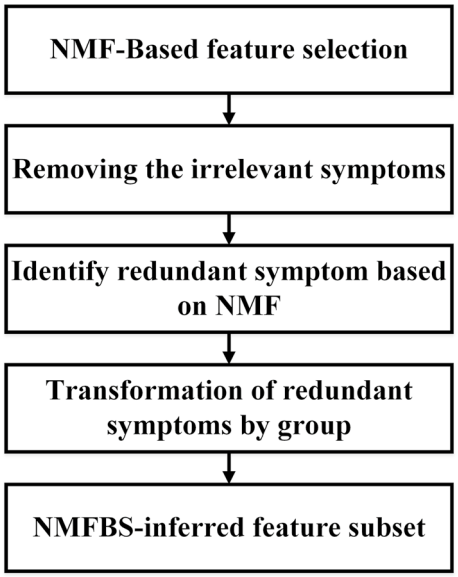

Fig. 7 Flowchart of NMF

problem, the shortest distance known as geodesic distance is introduced between the vertices which tend to be measured on a paired manifold and retain the local neighborhood points (Lee et al. 2004). However, Isomap algorithm determines the low dimensional space vector, which is capable to preserve the geodesic distance between two features. These features are determined from the sampled dataset and scoped along sub manifold. Figure 8 shows the algorithm of Isomap.

Figure 8 depicts the algorithm of Isomap in which the first steps is, a neighborhood graph is built on the manifold. The second step is to find Euclidian and geodesic distance between two features. The last step is to form the low dimensional data set from HDD. To explain the Isomap algorithm briefly, Fig. 9 shows an example of Swiss roll dataset between two samples *A*&*B*.

Figure 9a depicts the Euclidian distance between samples *A* and *B* by applying MDS whereas Fig. 9b gives the geodesic distance between *A* and *B*. The colour codes represent the number of class matrices. In general, Isomap is proved to be a faithful algorithm to represent the HDD globally. However, if the data are not sampled uniformly, Isomap may create topological instabilities as this algorithm doesn't stand for any isometric embedding for global graphical representation. Hence, it must be necessary to flourish the range of manifold whose Euclidian distance must be closer to local geometry of HDD instead of its global geometry.

2.5.2 Locally linear embedding (LLE) algorithm

Distinction to Isomap, LLE method holds the local neighborhood points. It predicts that each original data points are combined linearly to their nearest neighbors, and tries to be converted into lower dimension data points to maintain linear combination characteristics. Figure 10 shows the algorithm of LLE.

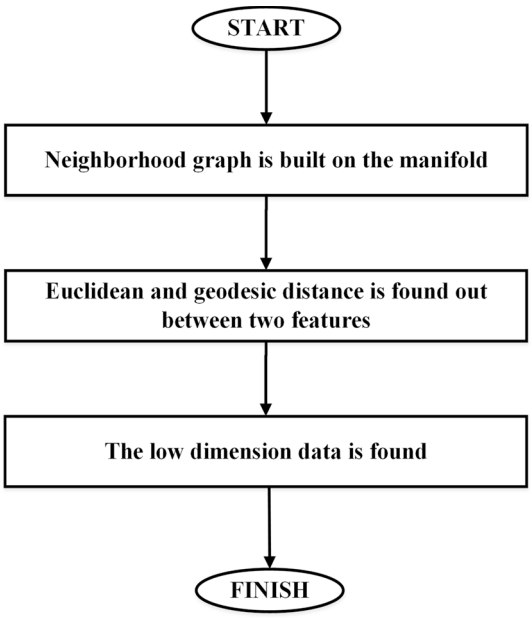

Fig. 8 Algorithm of Isomap

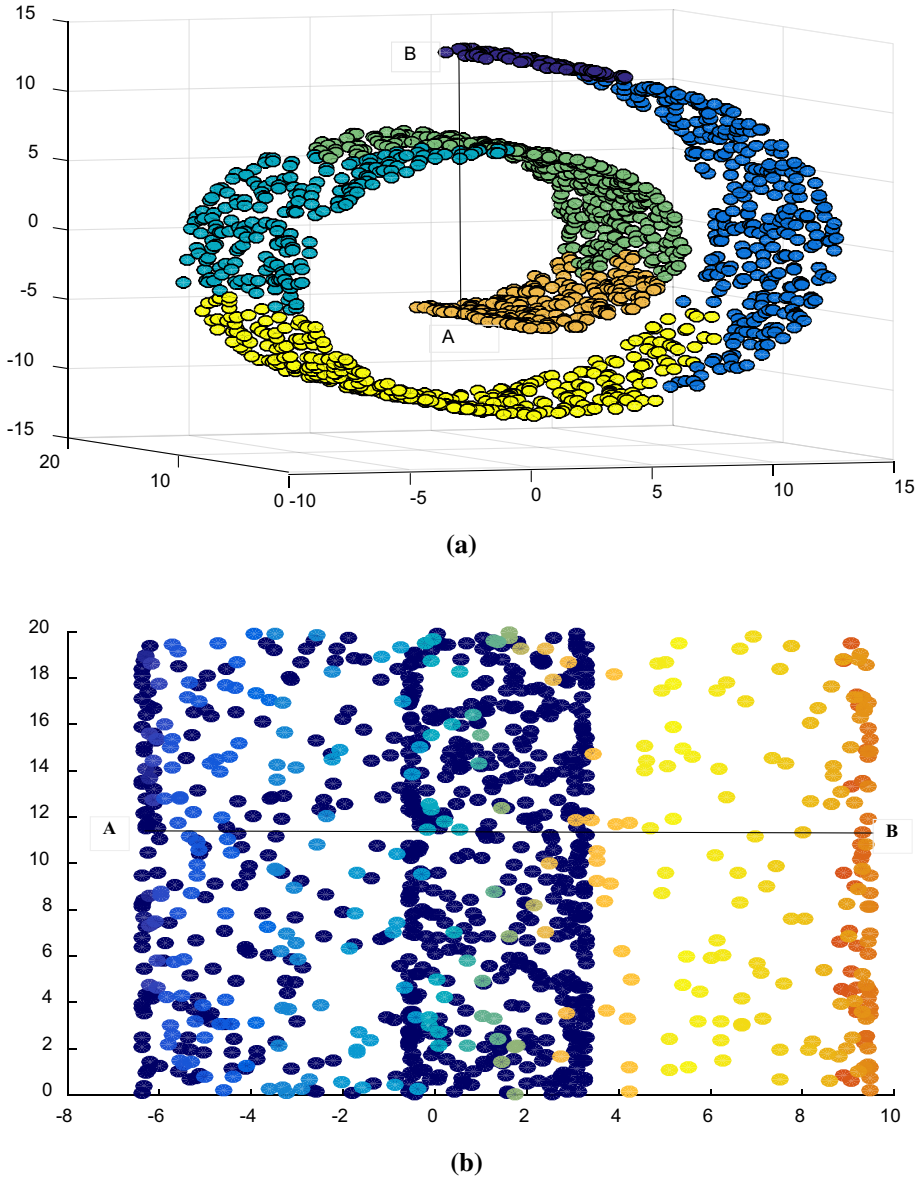

Fig. 9 Swiss Roll dataset with **a** Euclidian distance **b** geodesic distance

Figure 10 depicts the algorithm of LLE in which data point a_i, which is represented as the linear combination of α_{ij} to its k neighboring point of a_j. A dissimilarity matrix Δ is ranked to identify the k nearest point. The expression for LLE is given as:

$$\left\{ \min_{x} \sum_{i=1}^{p} \left\| X_i - \sum_{j=1}^{k} \alpha_{ij} X_{ij} \right\|^2 \right. \tag{17}$$

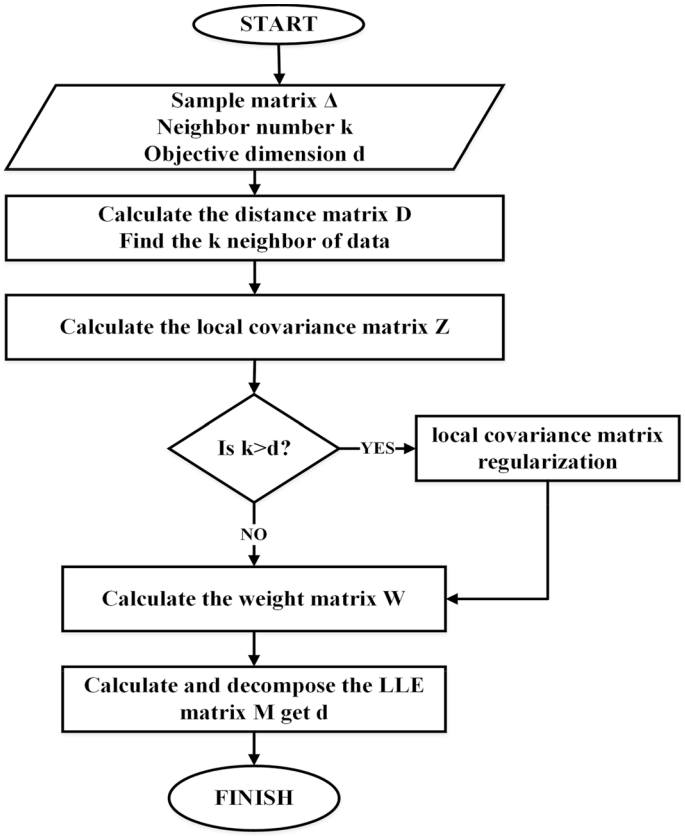

Fig. 10 Algorithm of LLE

where $\frac{1}{p}X^T X = I$. In Eq. (17), X is the solution matrix whose kth column is indicated by x^k and p is the number of data points for the dataset X. The trivial solution is avoided for the constraint $X = 0$. The dissimilarity matrix Δ can be modified into Δ^*, such that $\Delta^* = \Delta + \omega\max(\Delta)$, the corresponding LLE method can be fully supervised. It is noted that $\max(\Delta)$ is the maximal Δ entry. LLE algorithm is computationally efficient because of the sparsity of neighborhood data points. Hence, it is helpful in preserving the local geometry of data by mapping the neighborhood points on the manifold to represent low dimensional data (Chen and Yang 2011; Sorzano et al. 2014).

2.5.3 Laplacian eigenmap (LE) algorithm

By using Laplacian eigen maps of data mapping, LE algorithm tries to preserve the locality of structure. The similarity matrix can be formed by using Gaussian function such as:

$$\alpha_{ij} = \exp\left(\frac{-||X_i - X_j||^2}{\gamma}\right) \tag{18}$$

where γ is the scaling parameter which is the mean of the squared distances of all paired points for $i,j = 1, 2, \ldots, p$. The lower dimensional data can be depicted by minimizing Eq. (18) as:

$$-\frac{1}{2}\sum_{i,j}\left\|\mathrm{X}_i - \mathrm{X}_j\right\|^2 \alpha_{i,j} = tr(X^T L X) \tag{19}$$

Hence, LE is formulated as:

$$\begin{cases} \min\limits_{\mathrm{X}}(\mathrm{X}^T L X) \\ s.t. \mathrm{X}^T D \mathrm{X} = I \\ \mathrm{X}^T e = 0 \end{cases} \tag{20}$$

where I is known as identity matrix and e is represented as $(1, \ldots 1)^T$. The diagonal matrix has resemblance with dissimilarity matrix in terms of rows and columns. The Laplacian matrix is formulated by $\mathrm{L} = \mathrm{D} - \alpha$. Here, X is the required low-dimensional matrix. The constraints obtained from Eq. (20) helps in avoiding the trivial solutions at $\mathrm{X} = 0$ and $\mathrm{X} = e$ (Malik et al. 2016; Verónica et al. 2017). Isomap, LLE and LE algorithms are fit for non-linear data dimensional reduction on the basis of graphical approach and so susceptible to overcome the "curse of dimensionality" problem. Also the manifold requires a large number of data points to exploit its properties with intrinsic dimensionality but sometime the properties of local structured data don't follow the global structure which causes over fitting problems in manifold. Figure 11 shows the algorithm of LE.

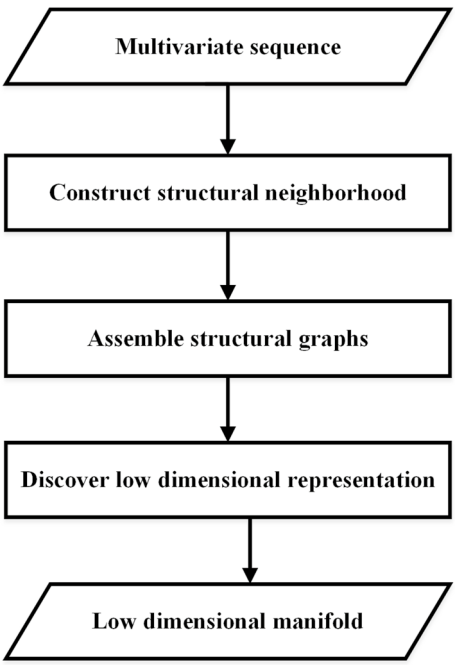

Fig. 11 Algorithm of LE

Figure 11 shows the algorithm of LE. The first step is to find multivariable sequence. Then the structural neighbourhood has been constructed. The next step is to assemble structural graphs. Then the low dimensional data has been found.

3 Feature selection algorithm (FSA)

In this section, the algorithms of feature selection have been discussed. The objective of any FSA is to select the best optimal feature set by eradicating the irrelevant features from the original dataset without any data transformation for the analysis of high dimensional data. Many researchers have explained their opinions in determining the best objective functions for variable selection of algorithms in several articles (Verónica et al. 2017; Ang et al. 2016).

On the basis of various methods of searching, FSAs are further branched as filter, wrapper and embedded methods. The most discriminative features are opted from the character of data by filter methods. Generally, filter methods perform feature selection before classification and usually involve a two-step process. In the first step, all the features are ranked according to a certain criteria. In the next step, the features with the highest ranking are selected. Many filter type methods have been studied in which most of them are univariate in considering the independent or dependent feature variables as per their ranking priorities. Wrapper methods use the intended learning algorithm to select a feature subset as search problem where various problems are prepared, evaluated and compared with other combinations. Embedded methods perform feature selection in the process of model construction (Das 2001). Table 1 depicts the attributes of FSA briefly.

In Table 1, filter type methods have been studied in which most of them are univariate in considering the independent or dependent feature variables as per their ranking priorities. Hierarchy of FSA has been shown in Fig. 12.

From Fig. 12, we observe that to distinguish various FSA, different feature-ranking methods are taken into account. The obtained results from feature ranking are highly dependent on individual features. The results of feature subset selection are also applicable to the target class. Further the results of embedded methods are prognostic models that can be configured by feature subsets. Filter methods include algorithms such as ReliefF, MRMR (Minimum Redundancy Maximum Relevance) etc. Wrappers are quite faster than filters, they use model hypothesis by taking the training data in focus. Feature dependency can be found out by wrapper methods. Hybridization of Genetic Algorithm (GA) and Particle Swarm Optimization (PSO) is one of the most popular algorithms used by wrapper type FSA.

Table 1 Attributes of different FSA methods

Types of FSA	Attributes	
	Classifier dependency	Computational cost
Filter	Independent	Low
Embedded	Dependent	Low
Wrapper	Dependent	High

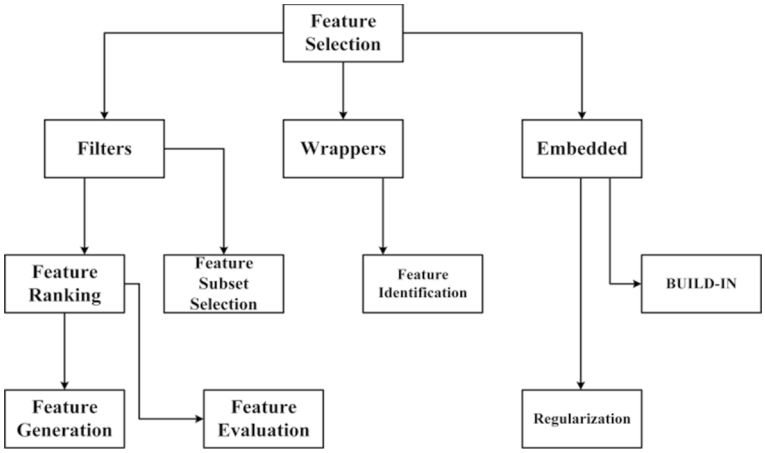

Fig. 12 Hierarchy of FSA

An effective feature-ranking algorithm alleviates the curse of dimensionality problems in large-scale data. Feature ranking otherwise known as variable ranking, a process of ordering the features by a scoring function that measures feature relevancy (Hsu et al. 2011). The score function is calculated from the training dataset. It results in better classification performance and a predictive generalized model is formed with lower computational cost which selects '*n*' significant features according to their priority in the model. The results are not generally optimal but for calculation purpose, it is computationally efficient. It involves algorithms such as GA, PSO, ReliefF, Minimum Redundancy Maximum Relevance (MRMR), Recursive Feature Elimination (RFE) and Simultaneous Perturbation Stochastic Approximation (SPSA).

3.1 Hybridised genetic algorithm and particle swarm optimization (HGAPSO)

Various schematic FSA demand for more training samples to evaluate the dataset accurately. In aid to that, the exhausted CPU processing time and computational time of samples is a prolonged issue for finding the "best" optimal features. To overcome this barrier, a faster FSA has been introduced known as HGAPSO. With an effective fitness function, this method has the ability to deal with HDD with minimum number of features. Here, the concept of two most popular optimization algorithms such as GA and PSO are recalled (Al-Rawi and Karajeh 2007).

3.1.1 Genetic algorithm (GA)

Inspired from the Darwin's theory of evolution, a novel computational approach hasbeen implementedknown as GA. It is a random feature selection technique, which drives to a wide range of feasible solution. This algorithm engages a population of parameters where each parameter is selected for a minimum fitness function and forms a unique generation to evaluate the probability of selection (Pedram and Benediktsson 2015). The algorithm of GA is shown in Fig. 13.

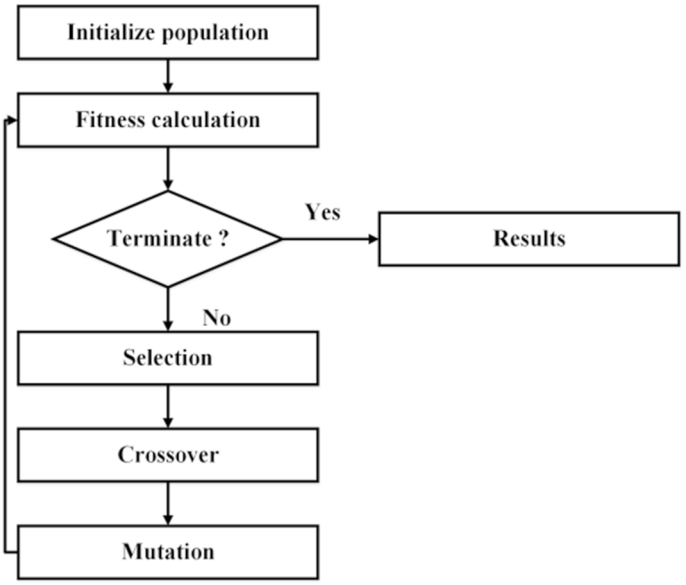

Fig. 13 Flowchart of GA algorithm

Figure 13 shows the algorithm of GA. Feature selection by GA involves three major steps to accomplish i.e. Reproduction, Crossover and Mutation. Here, in this algorithm, linear ranking method is used for identifying suitable genes from a set of chromosomes. In the ranking method all the chromosome's in a matrix is assigned some rank based on the fitness function. Further based on the highest rank, the chromosome'sare selected. The fitness function may be considered as the way to define the resultant accuracy. The detail procedure of GA based feature selection is reported in Pedram and Benediktsson (2015).

3.1.2 Particle swarm optimization (PSO)

PSO was first derived by Kennedy and Eberhart as a biologically inspired technique, observing the flocking sound of birds (Mahale and Chavan 2012). Unlike GA, the population consists of a set of particles where each particle is the combination of parameters. The word "Swarm" consists of a set of particles in a multi-dimensional space, each group travels with a unique velocity. The algorithm of PSO is given in Fig. 14.

Figure 14, shows the algorithm of PSO, where the current and preceding positions of each particle can be tracked by a memory chip, which is classified as "personal best positions" and "global best positions". The non-causal behaviour of a preceding particle and its neighbor particle is observed to change the velocity of current particle in the search space. A particle is highly dependent on three factors: (a) Current position of the particle, (b) memory required to observe the preceding position of the particle, (c) The knowledge of swarm about the particle set. Hence, the particles tend to move toward the search space where they can travel with a desired velocity (Mahale and Chavan 2012).

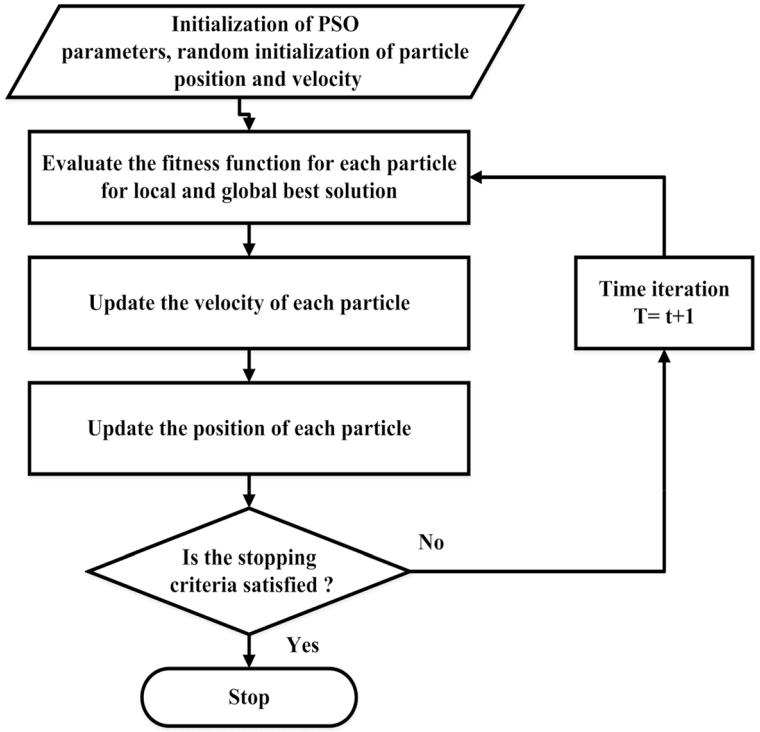

Fig. 14 Flowchart of PSO

3.1.3 Hgapso

GA can be incorporated with PSO by integrating the steps such as GA based selection, crossover and mutation with an updated position and velocity of particles from PSO (Hong and Dong 2004). Hence, HGAPSO is an integrated technique used to select the features optimally from a set of population in which the number of features must be equal with total dimension of particles. In this technique, the dimension of position is equal with the dimension of velocity of each particle as:

$$v_i^k = x_i^k = r \tag{21}$$

where x_i^k and v_i^k is the dimension of position and velocity of a particle respectively. The positions are taken in terms of binary values of 0's and 1's representing absence and presence of features respectively (Tao et al. 2018). Figure 15 shows the flowchart of HGAPSO for feature selection.

In Fig. 15, a set of random population is rendered and by applying optimization algorithm fitness values are acquired for each parameter of the population. Then by sorting method, the features are placed in descending order. After that, combinations of GA and PSO are implemented where the individual population is carried out as chromosomes for GA and as particles for PSO. Then the whole set is divided into two different subsets, each retaining its own attributes.

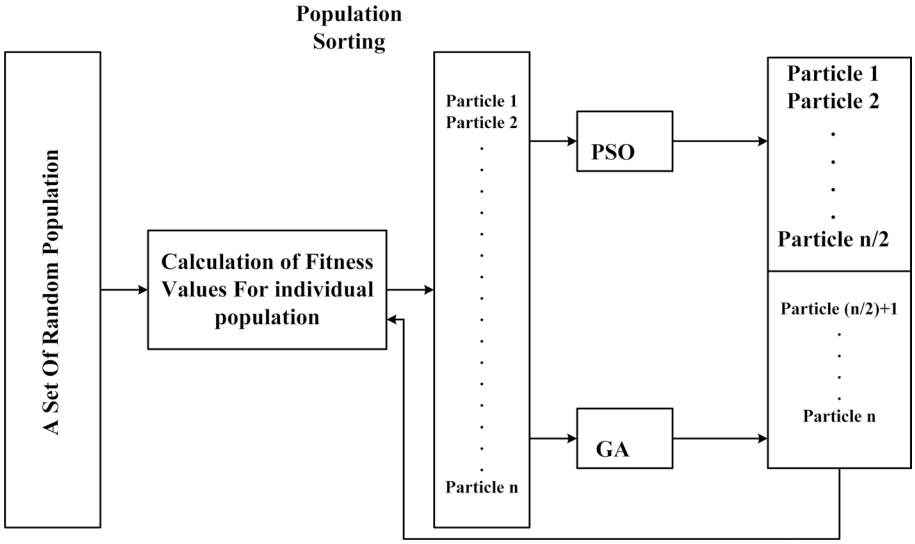

Fig. 15 Flowchart of HGAPSO

3.2 ReliefF

Filter-method approach has been taken for ReliefF algorithm. Despite the incomplete data, ReliefF selects the feature depending on the quality attributes estimation. ReliefF algorithm is a popular technique for feature weight computation. This method can be utilized in nominal, numeric, incomplete and noisy data. Generally, filter methods are faster than wrapper method. It is a unique family of filter type feature selection algorithm that maintains a satisfactory balance between the objective and computational accuracy. Feature relevance defines feature weight as a measurement vector by solving an optimization problem. However, there are some shortcomings of this algorithm such as frequency sampling is uncertain and there is a fluctuation in feature weight instances. Hence, some modification in ReliefF algorithm is needed to resolve the accuracy related problem on the basis of mean variance model. Feature weights are estimated by considering the mean and variance of discriminative sample data to attain more accurate results (Zhi et al. 2016).

The input to find the optimal features is a feature weight vector $F_w(A)$ for class C and for other attribute values. The output is aimed at determining the qualities of attributes for the input raw signal. To achieve this output, some steps are followed which is shown in Fig. 16.

Figure 16 depicts the algorithm of ReliefF. In step 1, a weight F_w of sample data matrix is fixed at 0. Then, the process begins for $i = 1, \ldots, n$. For an instance r_i is randomly chosen and class (r_i) is computed for each class C. In the next step, the hits h_j points and misses the point $n_j(c)$ of k nearest points are determined for each class C. Again, the weight vector is computed as,

$$F_w(A) = F_w(A) - (a_i - h_i)^2 + (a_i - n_j(c))^2 \tag{22}$$

Then, the entire procedure is repeated for n iteration and each weight vector is divided by n to get a relevance vector. To verify the feature weight vector, a threshold value T_h is taken and compared with relevance vector for optimal selection of features.

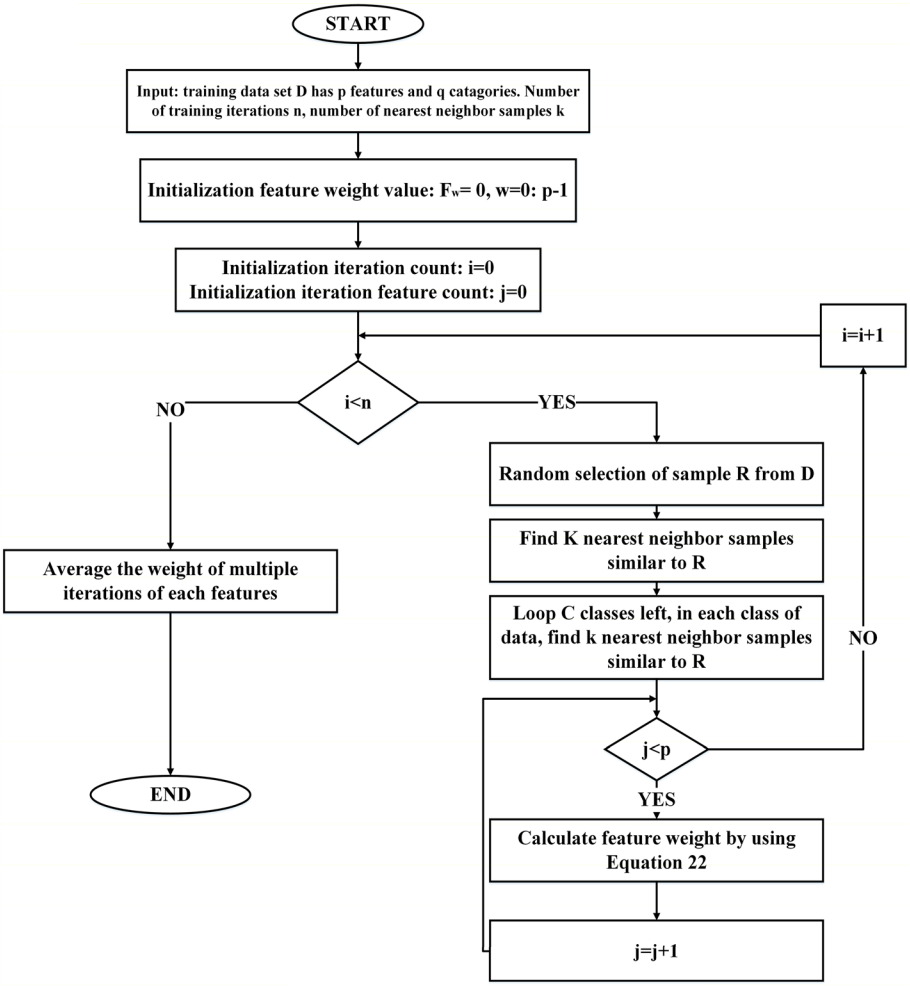

Fig. 16 Algorithm of ReliefF

The detail algorithm can be studied from Ghosh et al. (2013). Figure 9 gives a graphical representation for two different classes such as class 'O' and class 'X' showing how features are selected optimally.

From Fig. 16, it is noticed that ReliefF algorithm is not bounded with feature selection of a single class and deals with multi-classes. The target and neighbor instance and target instance is searched for the same class which is also known as h_j points. Here also misspoints $n_j(c)$ are found out for the search of k nearest points. The class distribution is shown in Fig. 17.

In Fig. 17, the selection of instances occurs randomly due to which the sampling frequency becomes uncertain and hence this algorithm reduces the accuracy. The descriptions of all the instances are shown in Table 2.

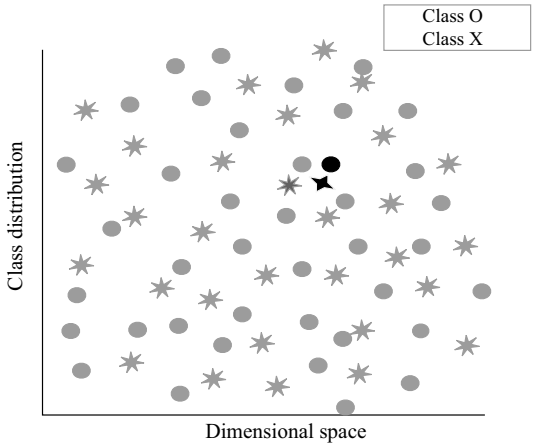

Fig. 17 Optimal feature selection by ReliefF

Table 2 depicts the state-of-art of ReliefF algorithm which is more effective than any other wrapper methods by calculating the feature weight vector and also figures out convex optimization problems.

3.3 Minimum redundancy maximum relevance (MRMR)

Before expecting any feasible solution, the strength of feature relevance should be taken into account. Strong feature relevance posses the removal of irrelevant features based on predictive accuracy, where as weak feature relevance indicates the contribution of features to predictive accuracy. Along with feature relevance, feature redundancy is somehow intersected in FSA. Feature redundancy can be well defined in terms of two perfectly correlated features as well as two independent features (Naji et al. 2019; Zhou et al. 2019; Wang et al. 2018a, b; Jinxing et al. 2017). In a more common way, feature relevance and feature redundancy are dependent on a feature subset. To alter this problem, identification ofthe redundant features or removal of the non-redundant features without affecting the dimension of original feature set is required.

Mutual information is the most powerful tool in finding and selecting relevant features from a feature space. Often, some of these subsets contain some characteristics of the features, which may be redundant. Since our aim is to select the most relevant feature subset with a limited size factor, a new approach such as MRMR selection algorithm has been developed (Zhou et al. 2019; Wang et al. 2018a, b; Jinxing et al. 2017) to eradicate this problem. Let us consider two random variables A and B such that their mutual information I_M can be defined as the measurement of mutual dependence such as:

$$I_M = \sum_{a \in A} \sum_{b \in B} p(a,b) \log \left[\frac{p(a,b)}{p(a)p(b)} \right] \tag{23}$$

where $p(a)$ is the marginal probability of A and $p(a,b)$ is the joint probability of A and B. The selections of optimal features involve 2 steps: (a) Measurement of the dependency between the features to be selected and the target dataset by mutual information (b) Maximization of the dependency. Since, the calculation of dependency is seemed to be a tough job; hence we choose to find out the dependency by MRMR criteria as per the formula:

Table 2 Description of instances

Instances	Attributes
	Target instance for class O
	Nearest neighbor instances forclass O
	Nearest neighbor instances forclass X
	Zero weight instances forclass O
	Zero weight instances forclass X

$$\text{dependency} = \text{relevance - redundancy} = I_M(\mathrm{a}_j;\mathrm{c}) - \frac{1}{n-1}\sum\nolimits_{\mathrm{a}_i \in \mathrm{S}_{n-1}} I_M(\mathrm{a}_j;\mathrm{a}_i) \quad (24)$$

where a_i is the total number of feature subset and a_j is the features chosen for consideration under the set $\mathrm{A} - \mathrm{S}_{n-1}$, where S_{n-1} is selected feature subset. Let S_t represents the feature set to be selected and δ be the classification target. Hence, we have to find out the correlation between a feature, a_t of S_t and δ by the formula:

$$\bar{C} = \mathrm{I}_{\mathrm{M}}(a_{\mathrm{t}}, \mathrm{n}) \quad (25)$$

Redundancy between S_{n-1} and S_t can be represented as;

$$R_e = \frac{1}{n}\sum\nolimits_{\mathrm{a}_i \in \mathrm{S}_{n-1}} I_M(\mathrm{a}_t, \mathrm{a}_i) \quad (26)$$

The proper defined MRMR function can be obtained by combining Eq. (24) and Eq. (25) as:

$$\max_{\mathrm{a}_j \in \mathrm{S}_{n-1}} \left[I_M(\mathrm{a}_j, \delta) - \frac{1}{n}\sum_{\mathrm{a}_i \in \mathrm{S}_{n-1}} I_M(\mathrm{a}_j, \mathrm{a}_i) \right] \quad (27)$$

Equation (26) justifies the maximum correlation with minimum redundancy among the a_j features in S_t. In this algorithm, relevance is computed by using F-static values and mutual information for continuous and discrete features respectively whereas, redundancy is computed by applying Pearson and mutual information (Jinxing et al. 2017). The algorithm of MRMR is given in Fig. 18.

Figure 18, shows that data set has been divided into training set and validation set. Then the value of training set goes to MRMR and feature subset is selected. The feature subset and validation set have been passed through the quality estimation and selected feature has been passed through stopping criterion. After checking, the final feature subset is selected for the further process.

3.4 Recursive feature elimination (RFE)

RFE is a technique, which contains the weights of the feature vectors as an important attribute for optimal feature subset prediction. A certain fraction of features are eradicated after eachiteration based on the ranking priority. It must be noted that the features can predict the best result even though some of the feature cannot predict well in target data. Whendealed with noisy data, RFE-SVM algorithm can be useful in extracting sufficient information to respective weight values. The iteration process continues till the prediction errors are settled to a minimum value or specified feature values are eliminated. In Gysels et al. (2005), RFE conjoint with Support Vector Machine (SVM) is discussed for the analysis of a large dimensional dataset and the result accuracy is found to be accurate compared to any other methods. SVM was designed by Vapnik in 1963 to reduce the prediction error of optimally selected features. SVM plays a vital role in enhancing the accuracy of FSA and eliminate the non-redundant features in a recursive technique. Also RFE based SVM can be represented as a backward selection algorithm to solve the basic optimization problem. A soft margin SVM with linear Kernel function improves the quality of performance when we deal with massive number of features. The detail algorithm is reported in Sun et al. (2013). Figure 19 shows the algorithm of RFE.

In Fig. 19, the algorithm of RFE has been explained step by step,

1. Train model on all the features.
2. Repeat until tuning set accuracy decreases:
   - Rank features by coefficient magnitude in last model
   - Remove lowest-ranked features, e.g. bottom 10%
   - Retrain model
3. Return previous model (prior to accuracy decrease)

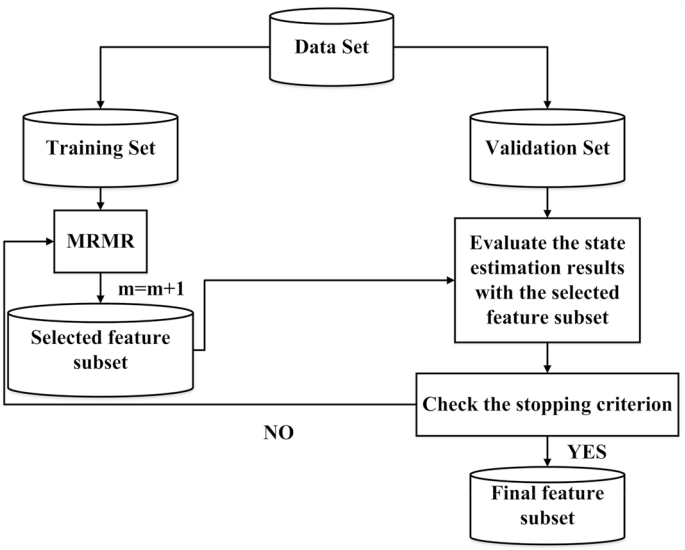

Fig. 18 Flowchart of MRMR

RFE is a backward selection wrapper method used with linear SVMs, but works for other linear methods too. It requires nontrivial modifications for linear methods, since those do not yield coefficient on features.

3.5 Simultaneous perturbation stochastic approximation (SPSA)

Often it is observed that gradient effects are incapable to compute the selected features because the objective functions are not optimized properly. So to avoid this problem, a gradient approximation method is used known as SPSA. This method is proved to be effective as it needs only two objective functions for stochastic gradient approximation (Zeren et al. 2018).

Let us consider an objective function $S{:}Z^{d} \rightarrow Z$ such that v^{*} is an optimum of S that fulfill the characteristics of gradient space, which disappears at point. To compute the gradient effectively, SPSA method is introduced which incorporates $\hat{g}(v^{*})$ which is considered to be an approximate for true value of $g(v^{*})$. Objective function performs the gradient descent in a stepwise manner such as:

$$\nu_{m+1} = \nu_m + \mathrm{d}_m \hat{g}_m(\nu_m) \tag{28}$$

where $\mathrm{d}_m \in Z$ is the size of the optimization step for each m iteration. For a perturbation vector $\mathfrak{R}_m \in Z^{d}$, the gradient approximation is calculated for the given $\mathfrak{R}_m$ as:

$$\hat{\mathrm{g}}_m(\nu_m) = \begin{bmatrix} \frac{1}{\mathfrak{R}_{m1}} \\ \frac{1}{\mathfrak{R}_{m2}} \\ \vdots \\ \frac{1}{\mathfrak{R}_{md}} \end{bmatrix} . \frac{\mathrm{S}(\nu_m + \mathrm{C}_m \mathfrak{R}_m) - \mathrm{S}(\nu_m - \mathrm{C}_m \mathfrak{R}_m)}{2\mathrm{C}_m} \tag{29}$$

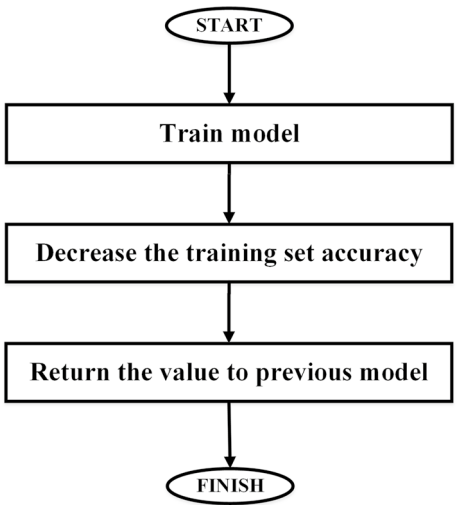

Fig. 19 Algorithm of RFE

where $\mathfrak{R}_m = \{\mathfrak{R}_{m1}, \mathfrak{R}_{m2} \ldots, \mathfrak{R}_{md}\}$ is a vector of δ and the elements are mutually independent mean-zero random variables. The satisfying criteria for two finite constants α_0 and α_1 are $|\mathfrak{R}_{ml}| < \alpha_0$ and $E|\mathfrak{R}_{ml}^{-1}| < \alpha_1$ respectively. When we deal with extreme non-linear and noisy data, multiple numbers of SPSA gradients can be calculated and their mean can be accomplished for each iteration (Zeren et al. 2018; Mathias et al. 2018; Inbarani et al. 2014). Figure 20 shows the algorithm of SPSA.

The detail algorithm of SPSA can be studied from Fig. 20 which depicts the random selection of a constant finite coefficient α to obtain two finite constants such as α_0 and α_1. After simulating both the constants, the satisfying criterions have also been calculated. Then the optimality condition has been checked. In case of non-feasible solutions, the gradient has been searched and values of α has been updated for each iteration (Zeren et al. 2018).

Tables 3 and 4 show the advantage and disadvantage of FEA and FSA.

4 Comparison analysis

A comparison analysis has the prospect of generalizing valuable understandings between various DRTs based on their qualitative merits and some of the demerit key factors (Fu 2016). For a brief observance at a glance, advantages and disadvantages of all the DRTs are mentioned in Table 3 and it can be noticed that transformation of original features into a new feature subset may create the data loss and data complexity problem. To avoid this, FSAs are implemented which makes the training model less complicated for further processing. Hence, the state-of-art of FSA is helpful not only for resolving the over fitting issues but also insights new approaches to overcome the "Curse of dimensionality" (Hira and Gillies 2015; Jothi and Rajam 2017).

FEA methods such as PCA, LFDA, CCA, NMF and manifold based algorithms are listed out in Table 4. Hence, the state-of-art of FSA is helpful with several FSAs such as HGAPSO, ReliefF, MRMR, REF and SPSA as illustrated in the same Table 4.

The healthcare data has been taken for comparison analysis and different dimension reduction techniques have been used to reduce the dimensionality for prominent classification. The different methods of feature extraction discussed here are PCA, LFDA, CCA, NMF and Manifold based algorithms. To figure out its significance, advantages and disadvantages, different feature selection methods are segregated into filter, wrapper and embedded type techniques such as HGAPSO, ReliefF, RFE-SVM, MRMR and SPSA. However, feature extraction differs from feature selection in terms of maintaining the originality of the signal and promotes the transformation of original feature subset into a new feature subset with uncertain data loss. But feature selection scheme chooses feature optimally without disturbing the attributes of original dataset. Also during the extraction process, some features may be transformed into new features, which may lead to overfitting problem and a miserable chance of data loss. To overcome these problems, FSA based schemes are employed, which not only solve the curse of dimensionality but also opt the most redundant features with accurate feasibility.

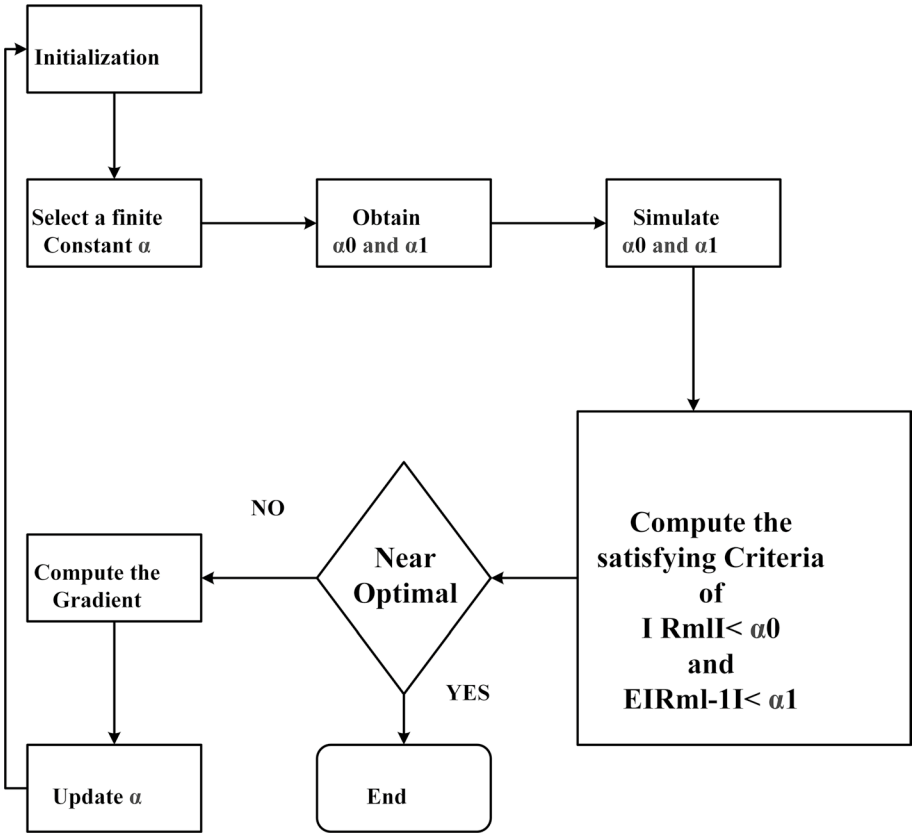

Fig. 20 Algorithm of SPSA

Table 3 Advantages and disadvantages of feature extraction and feature selection methods

Methods	Advantages	Disadvantages
FEA	Makes the domain of signal processing and image processing easier Quantifies the behavior of the features Discriminative power is more compared to any other data dimensional techniques Enhance the effectiveness of Machine Learning in supervised data dimensional reduction	Feature transformation is expensive There is a chance of data loss during new feature formation
FSA	Computation time to train the data model is faster It simplifies the model and makes it interpretable Choosing optimal features enhances the model accuracy It can overcome the over fitting problem	Optimal feature may not be feasible Selection of individual parameter is hard.

Table 4 Advantages and disadvantages of different DRTs

DRTs	Techniques used	Advantages	Disadvantages
FEA	PCA (Behbahani et al. 2017; Kapsoulis et al. 2018)	Fast and simple Minimum re-projection error Better reduction of noise	Optimal solution not feasible Data points are smaller than data dimension for which covariance matrix is large. Problem in calculation of covariance matrix Chance of data loss
	LFDA (Cong and Duan 2016; Li et al. 2015)	Computational ratio is less Accurate separation of large dataset for both with-in class and between class matrixes For feature transformation, the ranges of solutions are uniquely determined	Dependent on affinity matrix for choosing the features The probability interpretation lacks somewhere
	CCA (Jendoubi and Strimmer 2019)	Maximize the relationship between dependent and independent variables Achieve correlation with maximum accuracy.	Data loss occurs Results are poorly interrelated.
	NMF (Varghese et al. 2018)	Inspection of large data matrix is easy Common numerical approximation	Unable to show supervised data separation Poor convergence property
	Manifold (Zhang et al. 2010; Cannistraci et al. 2010)	Efficient algorithm for processing of data Mapping preserves properties of high dimensional data	Unable to handle noisy data Computation time is slower

Table 4 (continued)

DRTs	Techniques used	Advantages	Disadvantages
FSA	HGAPSO (Al-Rawi and Karajeh 2007)	Efficient selection of optimal features from a large dataset Implemented at its best where PCA cannot be used Finds probability of solution in less time.	Complexity in implementation. Choosing the appropriate parameters is hard Although the solution is metaheuristic, there is no guarantee of its accuracy
	ReliefF (Zhi et al. 2016)	Adaptive to all types of data Better computational efficiency Scalable in many iterative approaches.	Computational time is very expensive Fail to remove some of the redundant features.
	MRMR (Naji et al. 2019; Zhou et al. 2019; Wang et al. 2018a, b; Jinxing et al. 2017)	Provide more accurate solution Faster algorithm	Highly sensitive for parametric measurement
	RFE (Gysels et al. 2005)	Proper fitting of a model Continuously remove weak feature until specified features is achieved	Needs ML algorithm to be executed for every iteration Discriminative power
	SPSA (Zeren et al. 2018)	Efficient for the measurement of objective functionin high dimensional data Smooth optimized solution There is no over fitting issue in training the model Easy implementation	Sensitive to feature scaling Regularization of parameters for every iteration requires many hypermeters Some objective functions are non-smooth to be optimized

5 Validation for high dimensional data analysis

In the research area of signal processing, HDD analysis is predominating for more than 2 decades. To evaluate the complexity of the dimension of high dimensional data, we have considered some medical dataset measured from monitoring device and telemetryin two cases.

5.1 Description of data set

Microarray databases are a large source of genetic data, which upon proper analysis, could enhance our understanding of biology and medicine. Many microarray experiments have been designed to investigate the genetic mechanisms of cancer, and analytical approaches have been applied in order to classify different types of cancer or distinguish between cancerous and noncancerous tissue. The healthcare datasets of tumors mentioned in Tables 5 and 6 has high dimensionality. In this dataset, numbers of features are very high which has to be reduced for proper classification.

Case 1 The validation of proposed feature reduction algorithms can be verified by taking some real time dataset into account. Hence, three open source datasets has considered including three microarray datasets (CNS tumors, Leukemia and colon tumor). Table 5 summarizes the dataset with respective feature subset used for dimensional reduction algorithms.

In Table 5, the dataset instances vary from 60 to 93 for each disorder, number of features vary from 2000 to 6000 and class vary from 2 to 9. Rest of the information regarding the dataset is available in Alfaar et al. (2016), Stanojevic and Krivokapic (2014).

Case 2 The datasets used for this work are taken from repositories namely Kent-Ridge Biomedical Dataset Repository (Kent-Ridge 2005) and UCI (Lichman 2013). In this work, high-dimensional datasets as well as low-dimensional disease datasets are considered for the experiments. This work considers disease datasets on which feature selection techniques are applied to extract relevant and useful features. The details of the dataset are shown in Table 6.

Table 6 summarizes the dataset with respective feature subset used for dimensional reduction algorithms. In Table 6, the dataset instances vary from 60 to 104 for each disorder, the number of features vary from 2000 to 12600 and 2 classes have been determined. Rest of the information regarding the dataset is available in Jain and Singh (2018).

5.2 Experimental setup

To perform various feature extraction algorithms, the carcinoma datasets are segregated into training and testing data subset. The training and testing data do not merge to maintain the individual characteristics of each feature. The test model is formed from 70% of the training mode (Patro and Sahu 2015) and 30% of the testing mode for both the cases described below.

Table 5 Dataset description of case 1

Sl. no.	Datasets	Data instances	No. of features	Class
1	CNS tumours	60	5726	9
2	Leukemia	93	5327	3
3	Colon tumour	62	2000	2

Table 6 Dataset description of case 2

Sl. no.	Datasets	Data instances	No. of features	Class
1	Central nervous system	60	7129	2
2	Colon tumours	62	2001	2
3	Prostate cancer	104	12,600	2

Case 1 Here some of the features whose attributes are not prominent are taken as nominal values and is normalized within the range of [0, 1]. Most of the index features have been removed. Then, the complex & high dimensional data has been reduced into a lower dimensional matrix by the transformation of original feature set into an extracted feature subset by using some of the FEA and FSA. The FEA analyzed here are PCA, LFDA, CCA, NMF and manifold based techniques. The same procedures has been repeated for all the five proposed FSA such as HGAPSO, ReliefF, MRMR, RFE, and SPSA.

Case 2 Here the results are produced by applying the Relieff method on three chronic disease datasets. The proposed method works well with high dimensional microarray datasets. Table 7 shows the 'Number of Features Selected' corresponding to different thresholds for high-dimensional datasets. Threshold 'th1', 'th2' and 'th3' correspondingly represent mean, median and standard deviation of weights obtained from ReliefF method respectively.

5.3 Experimental outcome

Case 1 Three different high dimensional dataset including microarray data have been analyzed in order to reduce their dimensional size. For that, their necessary feature attributes need to be identified and reduced to a limited size. The numbers of features obtained from both FEA and FSA have been depicted in Tables 7 and 8 respectively.

From Table 7, it is observed that almost all the techniques have proved their efficiencies in extracting features from raw dataset. But, in NMF algorithm, the numbers of extracted features are less compared to others. Hence, the data processing becomes easier with less number of extracted features and simultaneously data complexity reduces. The reason behind the efficiency of NMF is the easy inspection of large data matrix without any data loss. From Table 8, it is concluded that the MRMR method reduces the redundancy of each features by filtering out the best one. MRMR method is the fastest algorithm to reduce the most redundant attributes and simplifies the feature subset matrix to perform an efficient classification performance. The graph between the number of features and feature

Table 7 Details of Features extraction by FEA

Datasets	FEA				
	PCA	LFDA	CCA	NMF	LE based
CNS Tumours	40	40	38	37	42
Leukemia	51	49	50	42	55
Colon tumour	16	16	18	15	21

Table 8 Details of feature selection by FSA

Dataset	FSA				
	GA	ReliefF	MRMR	RFE	SPSA
CNS Tumours	32	30	27	29	28
Leukemia	45	44	42	43	45
Colon tumour	11	8	7	12	10

extraction technique has been shown in Fig. 21 and the graph between the number of features and feature selection technique has been shown in Fig. 22.

However, comparing Figs. 21 and 22, it has been found that feature selection methods are implemented for removing irrelevant features efficiently as the no. of features chosen are reduced.

Case 2 Anew hybrid feature selection method based on ReliefF has been presented. The method is tested on three chronic disease datasets. For the selection of relevant & non-redundant features, three values of threshold are considered. Out of the three thresholds, an optimal value of threshold has been selected from the proposed approach. The whole feature selection process has been carried out. In Table 9, features selected for different dataset has been shown.

In Table 9, it has been shown that the ReliefF method generates a weight matrix & prepares a list of features by removing those with weights below the threshold value. The experimental results show that threshold 'Th3' performs best in terms of percentage of reduced features for high-dimensional datasets. The graph between the number of features and thresholds has been shown in Fig. 23.

In Fig. 23, it has been noticedthat the threshold 3 is showing better result than the other thresholds.

5.4 Classification

To evaluate the results more accurately, accuracy has been calculated for all the dimensional reduction techniques. For that, popular SVM classifier (David and Wien 2015), kNN, random forest, naïve bayes, decision tree and logistic regression classifiers (Suguna et al. 2019) have been taken. SVM is a direct learning method used in classification of text and images, bioinformatics. Here, the l_1 norm based SVM for linear regularization has been applied to perform the accuracy of various dimensional reduction techniques (Nie et al. 2010). KNN is non directive method. It is an ideal learning model with local estimation. Random forest is a group of model where many decision trees are grouped to make a

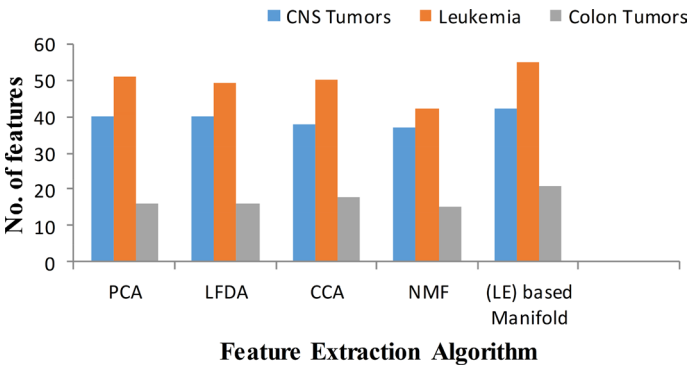

Fig. 21 Depicting the 'features extracted' for high-dimensional datasets

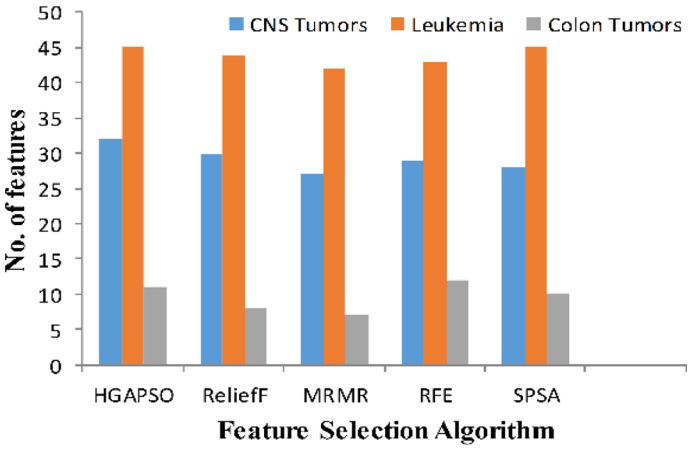

Fig. 22 Depicting the 'Features Selected' for high-dimensional datasets

robust model. It can handle overfitting better than any other model. Naive byes is a productive probability method used for classification problems. It is basically used for classification of text in which the set of feature is very large. Logical regression is a method where we put the feature and get the binary output using a logistic function.

Accuracy is the ratio between sample numbers of positive cases to the total number of samples. Accuracy shows how near the calculation to its positive value. F-score is the harmonic mean between precision and recall. F-score choose the preciseness of a classifier and it also tells the robustness of the classifier. Equations (30) and (31) shows the accuracy and f score calculation of SVM respectively.

$$\text{accuracy} = \frac{a + b}{a + b + c + d} \tag{30}$$

Table 9 Number of features selected by ReliefF

Sl. no.	Datasets	Total features	Threshold (th1)	Threshold (th2)	Threshold (th3)
1	Central Nervous System	7130	3262	3536	493
2	Colon tumour	2001	812	1000	264
3	Prostate cancer	12,600	6333	6300	35

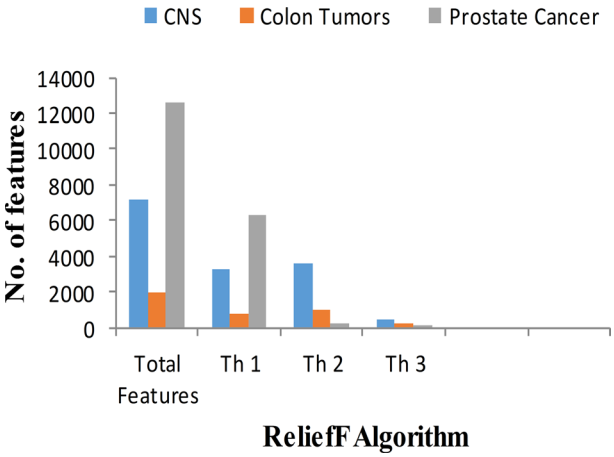

Fig. 23 Depicting the 'Features Selected' for high-dimensional datasets

$$\text{F score} = \frac{2 * b}{2 * b + c + d} \tag{31}$$

where Eqs. (30) and (31) show 'a' is true negatives (number of negative instances that is correctly classified as negative classes), 'b' is true positive (number of positive instances that is correctly classified as positive classes), 'c' is false negative (number of instances that is incorrectly classified as negative classes), 'd' is false positive (number of instances that is incorrectly classified as positive classes). Tables 10 and 11 shows the performance matrix of classifiers using FEA and FSA respectively.

Case 1 To evaluate the performance metrics of FEA and FSA by using different classifiers.

In Tables 10 and 11, it shows the accuracy (A) and F score (Fscr.) values of classifiers such as RF (Random Forest), DT (Decision Tree), kNN (k- nearest neighbor), LR (Logistic Regression), NB (Naive Bayes), SVM (Support Vector Machine). It has been observed that the SVM is classifying properly and is having a good accuracy value as compared to others.

Comparing Tables 10 and 11, it can be clearly observed that SVM plays a vital role in the classification of high dimensional data relevantly than any other classifiers. It is

better to choose a novel classifier. Figure 24 shows the ROC curve of NMF using SVM and Fig. 25 shows the ROC curve of MRMR using SVM.

In Figs. 24 and 25 ROC curve of NMF and MRMR using different classifiers has been shown respectively. The x-axis of graph is shown as the false positive rate and y-axis is shown as the true positive rate. Comparing the entire ROC curve, it has been found that the SVM classifier is doing better than any other classifier. Figure 26a, b display the accuracy of SVM of all the proposed FEA and FSA based DRTs respectively.

Comparing Fig. 26a, b, it can be clearly observed that FSA plays a vital role in the reduction of high dimensional data relevantly than any other FEA. It is better to choose a novel FSA for two reasons. First reason is that FSA can reduce the higher dimension of high dimensional data into its lower dimensional form effectively and improve the performance accuracy of the classifier. Secondly, it reduces the overfitting problem by training the dataset at the time of feature selection procedure itself; hence the computational time of training data is also trimmed.

Table 10 Performance metrics of classifiers using FEA

Data set	Classifiers	FEA									
		PCA		LFDA		CCA		NMF		LE	
		A (%)	Fscr.	A (%)	Fscr.	A (%)	Fscr.	A (%)	Fscr.	A (%)	Fscr.
CNS tumours	RF	91	0.95	92	0.96	92	0.96	92	0.96	90	0.94
	DT	88	0.92	89	0.92	89	0.92	90	0.93	87	0.90
	KNN	85	0.90	96	0.90	96	0.90	96	0.91	94	0.88
	LR	93	0.95	94	0.95	94	0.95	94	0.96	92	0.93
	NB	84	0.89	86	0.89	86	0.89	86	0.90	84	0.87
	SVM	**91**	**0.95**	**94**	**0.97**	**93**	**0.97**	**96**	**0.98**	**91**	**0.95**
Leukemia	RF	91	0.95	92	0.96	92	0.96	92	0.96	90	0.94
	DT	88	0.92	89	0.92	89	0.92	90	0.93	87	0.90
	KNN	85	0.90	96	0.90	96	0.90	96	0.91	94	0.88
	LR	93	0.95	94	0.95	94	0.95	94	0.96	92	0.93
	NB	84	0.89	86	0.89	86	0.89	86	0.90	84	0.87
	SVM	**93**	**0.95**	**95**	**0.97**	**95**	**0.97**	**95**	**0.98**	**92**	**0.95**
Colon tumours	RF	91	0.95	92	0.96	92	0.96	92	0.96	90	0.94
	DT	88	0.92	89	0.92	89	0.92	90	0.93	87	0.90
	KNN	85	0.90	96	0.90	96	0.90	96	0.91	94	0.88
	LR	93	0.95	94	0.95	94	0.95	94	0.96	92	0.93
	NB	84	0.89	86	0.89	86	0.89	86	0.90	84	0.87
	SVM	**92**	**0.95**	**92**	**0.97**	**90**	**0.97**	**95**	**0.98**	**89**	**0.95**

Bold indicates significant values

Table 11 Performance metrics of classifiers using FSA

Data set	Classifiers	FSA									
		GA		ReliefF		MRMR		RFE		SPSA	
		A (%)	Fscr.	A (%)	Fscr.	A (%)	Fscr.	A (%)	Fscr.	A (%)	Fscr.
CNS tumours	RF	91	0.95	92	0.96	92	0.96	90	0.94	92	0.96
	DT	88	0.92	89	0.92	90	0.93	87	0.90	89	0.92
	KNN	85	0.90	96	0.90	96	0.91	94	0.88	96	0.90
	LR	93	0.95	94	0.95	94	0.96	92	0.93	94	0.95
	NB	84	0.89	86	0.89	86	0.90	84	0.87	86	0.89
	SVM	**91**	**0.96**	**93**	**0.97**	**96**	**0.98**	**91**	**0.95**	**94**	**0.97**
Leukemia	RF	91	0.95	92	0.96	92	0.96	90	0.94	92	0.96
	DT	88	0.92	89	0.92	90	0.93	87	0.90	89	0.92
	KNN	85	0.90	96	0.90	96	0.91	94	0.88	96	0.90
	LR	93	0.95	94	0.95	94	0.96	92	0.93	94	0.95
	NB	84	0.89	86	0.89	86	0.90	84	0.87	86	0.89
	SVM	**93**	**0.96**	**95**	**0.97**	**97**	**0.98**	**92**	**0.95**	**95**	**0.97**
Colon tumours	RF	91	0.95	92	0.96	92	0.96	90	0.94	92	0.96
	DT	88	0.92	89	0.92	90	0.93	87	0.90	89	0.92
	KNN	85	0.90	96	0.90	96	0.91	94	0.88	96	0.90
	LR	93	0.95	94	0.95	94	0.96	92	0.93	94	0.95
	NB	84	0.89	86	0.89	86	0.90	84	0.87	86	0.89
	SVM	**92**	**0.96**	**90**	**0.97**	**95**	**0.98**	**89**	**0.95**	**92**	**0.97**

Bold indicates significant values

Case 2 To evaluate the performance metrics of relief algorithm by using different classifiers.

Table 12, shows the accuracy (A) and F score (Fscr.) values of classifiers such as RF (Random Forest), DT (Decision Tree), kNN (k- nearest neighbor), LR (Logistic Regression), NB (Naive Bayes), SVM (Support Vector Machine). It has been observed that the SVM is classifying properly and is having a good accuracy value. Figure 27 shows the ROC curve of threshold3 using SVM.

In Fig. 27 ROC curve of Threshold 3 using different classifiers has been shown. The x-axis of graph is shown as the false positive rate and y-axis is shown as the true positive rate. Comparing the entire ROC curve it has been found that the SVM classifier is doing better than any other classifiers.

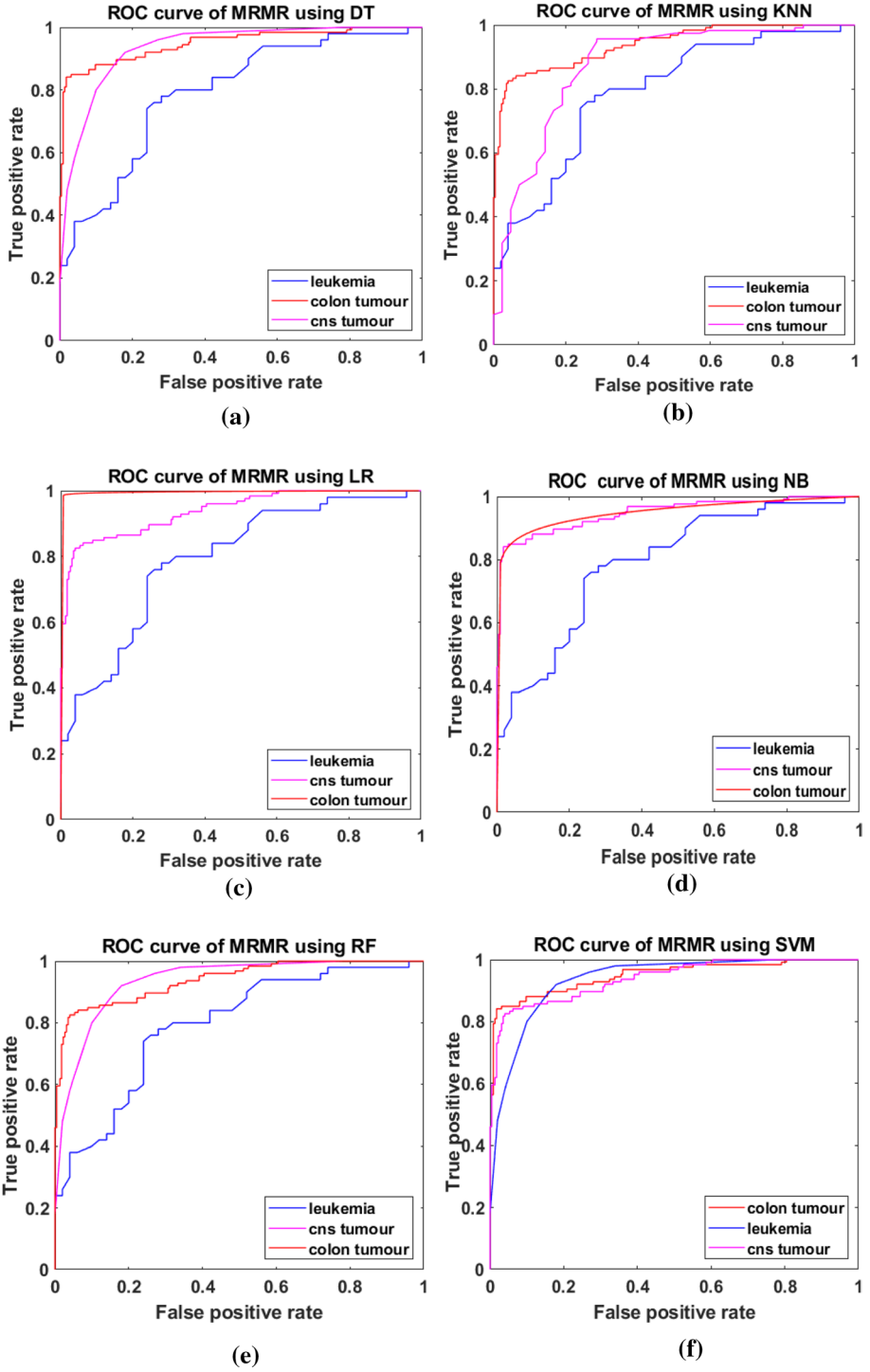

Fig. 24 ROC curve of NMF using different classifiers

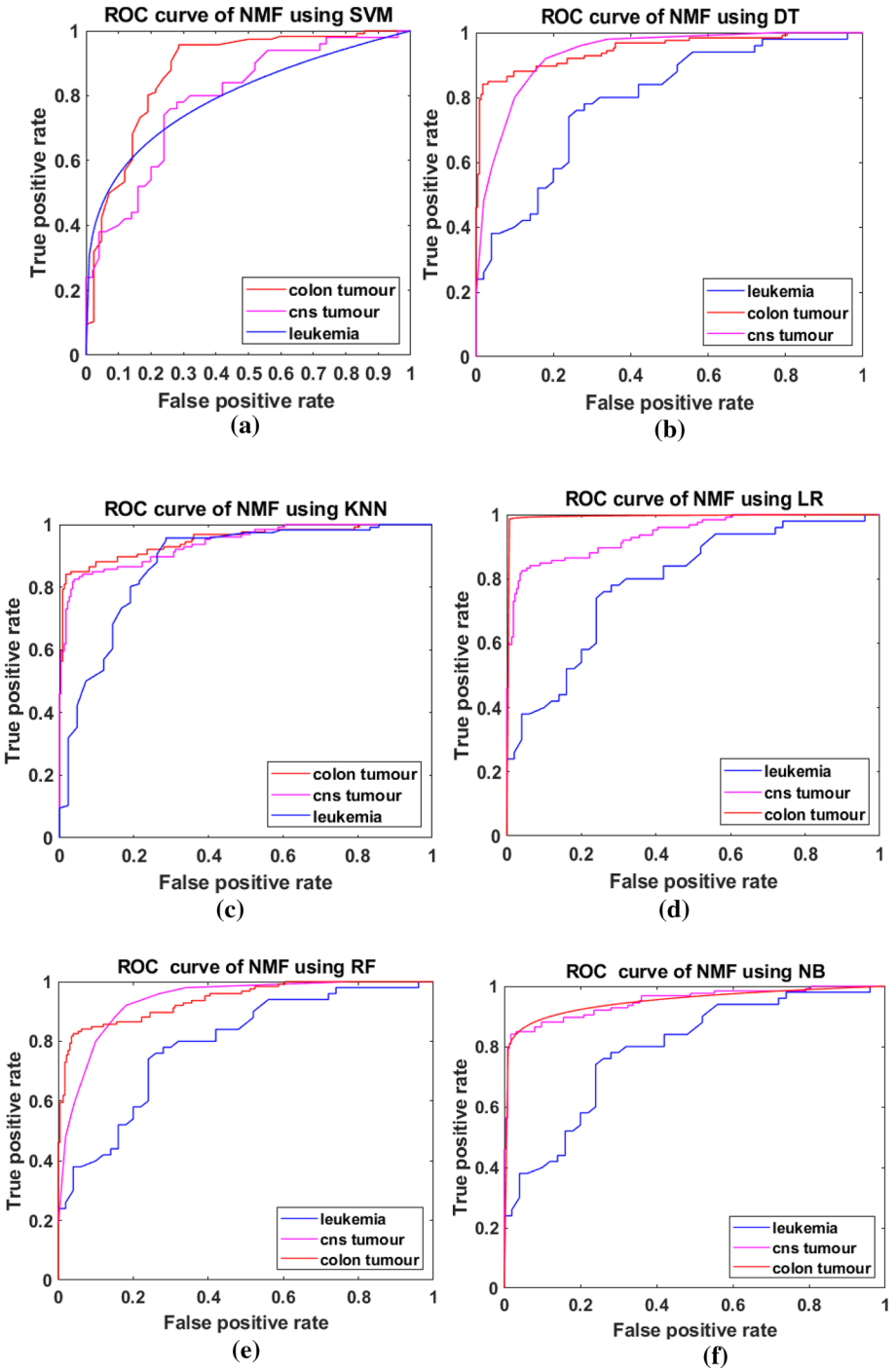

Fig. 25 ROC curve of MRMRusing different classifiers

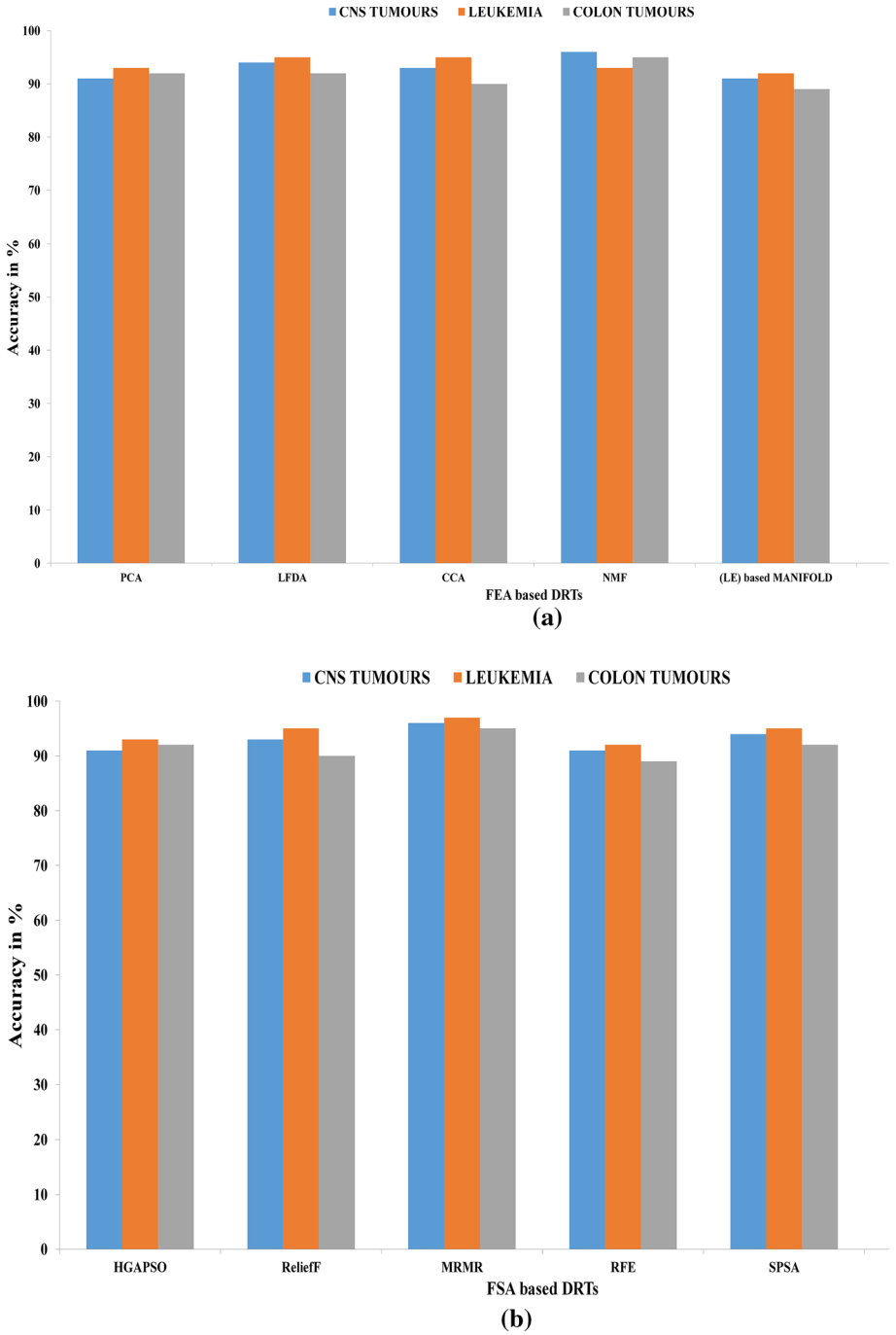

Fig. 26 Comparing accuracies of different FSA based DRTs **a** PCA, LFDA, CCA, NMF and LE based manifold **b** HGAPSO, ReliefF, MRMR, RFE and SPSA

Table 12 Performance metrics of classifiers using ReliefF

Data set	Classifiers	Threshold 1		Threshold 2		Threshold 3	
		A (%)	F scr.	A (%)	F scr.	A (%)	F scr.
CNS tumours	RF	90	0.94	92	0.96	92	0.96
	DT	87	0.90	89	0.92	90	0.93
	KNN	94	0.88	96	0.90	96	0.91
	LR	92	0.93	94	0.95	94	0.96
	NB	84	0.87	86	0.89	86	0.90
	SVM	**92**	**0.95**	**94**	**0.97**	**97**	**0.98**
Leukemia	RF	90	0.94	92	0.96	92	0.96
	DT	87	0.90	89	0.92	90	0.93
	KNN	94	0.88	96	0.90	96	0.91
	LR	92	0.93	94	0.95	94	0.96
	NB	84	0.87	86	0.89	86	0.90
	SVM	**92**	**0.95**	**94**	**0.97**	**97**	**0.98**
Prostate cancer	RF	90	0.94	92	0.96	92	0.96
	DT	87	0.90	89	0.92	90	0.93
	KNN	94	0.88	96	0.90	96	0.91
	LR	92	0.93	94	0.95	94	0.96
	NB	84	0.87	86	0.89	86	0.90
	SVM	**92**	**0.95**	**94**	**0.97**	**97**	**0.98**

Bold indicates significant values

6 Conclusion

This paper presents various methods of dimension reduction for high dimensional data briefly. Taxonomy of feature extraction method has been discussed and dividedinto PCA, LFDA, CCA algorithms on the basis of linear data analysis. Projection scheme to describe non-linear data by NMF and manifold based algorithms has also been considered. To figure out its significance, advantages and disadvantages, different feature selection methods are segregated into filter, wrapper and embedded type techniques such as HGAPSO, ReliefF, RFE-SVM, MRMR and SPSA. However, feature extraction differs from feature selection in terms of maintaining the originality of the signal and promotes the transformation of original feature subset into a new feature subset with uncertain data loss. But, feature selection chooses feature optimally without disturbing the attributes of original dataset. Comparing different FEA, it has been found that NMF based algorithm gives the accuracy in the range of 90% to 95%, which is highest among other feature extraction techniques. But, during the extraction process, some features may be transformed into new features, which may lead to overfitting problem and a miserable chance of data loss. To overcome these problems, it has been proposed to introduce various FSA, which not only solve the curse of dimensionality but also opt the most redundant features with accurate feasibility. The validation results indicate that MRMR method shows the accuracy in the range of 92% to 97%, which is comparatively better than all other proposed feature reduction techniques. Discussing many research articles, a conclusion is drawn that feature selection methods

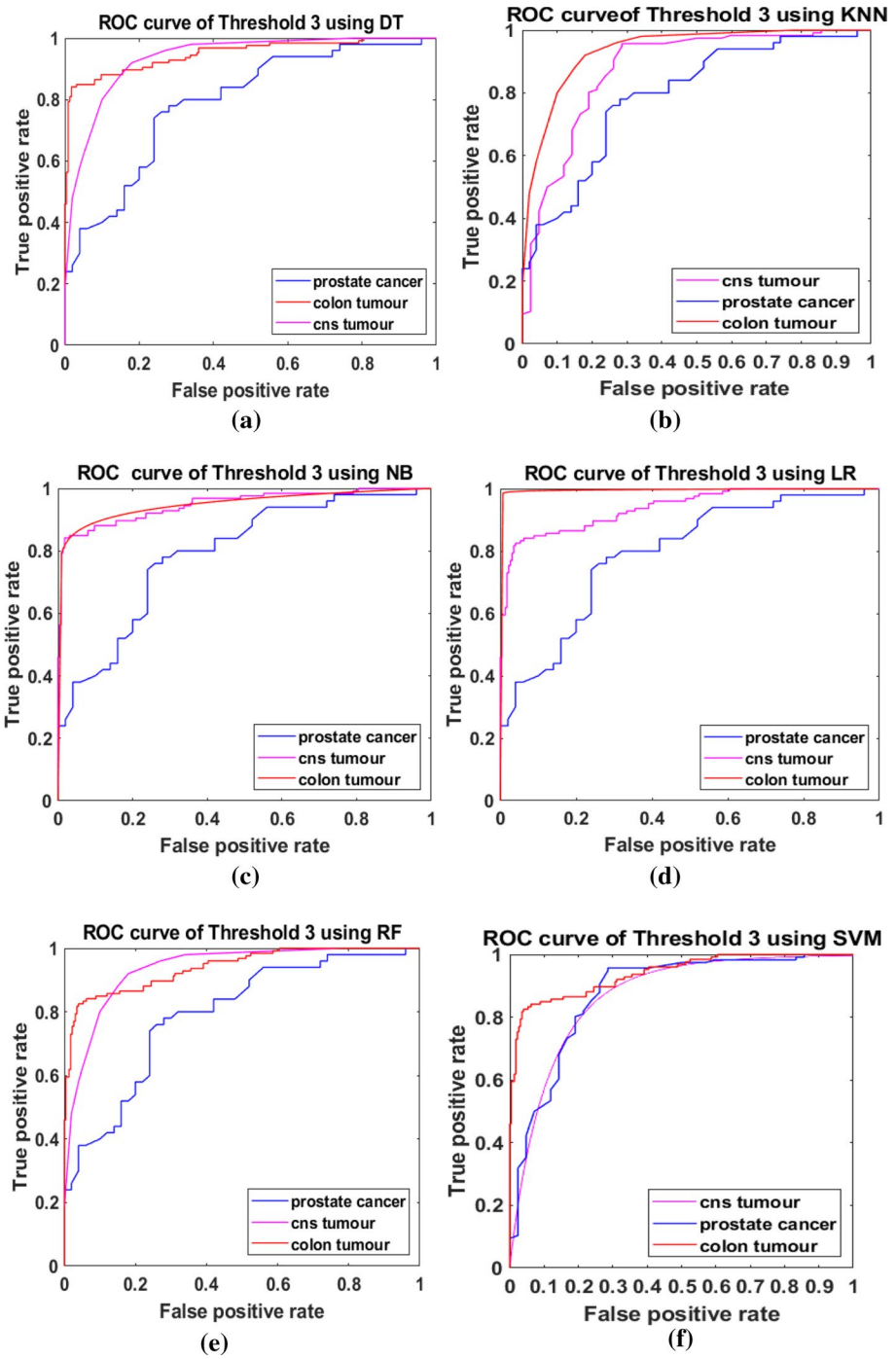

Fig. 27 ROC curve of Threshold3 using SVM

improve the pre-processing of data by removing its noise and reducing its dimensions so that a relevant feature subset can be opted for better classification performance.

References

Aggarwal CC, Cheng XZ (2012) Mining text data. Springer, Berlin

Al-Bakri NF, Soukaena HH (2018) Reducing data sparsity in recommender systems. Al-Naharin J Sci 21:138–147

Alexander CA, Wang L (2017) High dimensional data in healthcare: a new frontier inpersonalized medicine. Open Access J Trans Med Res 1–5

Alfaar AS, Waleed MH, Mohamed SB, Ibrahim Q (2016) Neonates with cancer and causes of death; lessons from 615 cases in the SEER databases. Cancer Med 6:1817–1826

Al-Rawi M, Karajeh H (2007) Genetic algorithm matched filter optimization for automated detection of blood vessels from digital retinal images. Comput Methods Prog Biomed 87(3):248–253

Ang JC, Andri M, Habibollah H, Haza Nuzly AH (2016) Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection. IEEE/ACM Trans Comput Biol Bioinf 13(5):971–989

Archenaa J, Mary Anita EA (2015) A survey of big data analytics in healthcare and government. Procedia Comput Sci 50:408–413

Behbahani BA, Yazdi FT, Shahidi F, Mortazavi SA, Mohebbi M (2017) Principle component analysis (PCA) for investigation of relationship between population dynamics of microbial pathogenesis, chemical and sensory characteristics in beef slices containing Tarragon essential oil. Microb Pathog 100(105):37–50

Cannistraci CV, Ravasi T, Montevecchi FM, Ideker T, Alessio M (2010) Nonlinear dimension reduction and clustering by minimum curvilinearity unfold neuropathic pain and tissue embryological classes. Bioinformatics 26(18):531–539

Chandrashekar G, Sahin F (2014) A survey on feature selection methods. Comput Electr Eng 40(1):16–28

Chen J, Yang L (2011) Locally linear embedding: a survey. Artif Intell Rev 36(1):29–48

Chen J, Zhang S (2009) Manifold learning based phoneme recognition. In: The proceedings of 2009 international conference on image analysis and signal processing, Taizhou, China

Cong I, Duan L (2016) Quantum discriminant analysis for dimensionality reduction and classification. New J Phys 18:1–10

Das S (2001) Filters, wrappers and a boosting-based hybrid for feature selection. In: Proceedings of 8th international conference on machine learning (ICML), vol 1, pp 74–81

David M, Wien FHT (2015).Support vector machines. The interface to LIBSVM in package, p 28

Deyan C, Zhao H (2012) Data security and privacy protection issues in cloud computing. In: Proceedings of 2012 international conference on computer science and electronics engineering, 1, pp 647–651

Ding S, Zhu H, Jia W, Su C (2012) A survey on feature extraction for pattern recognition. Artif Intell Rev 37(3):169–180

Fu MC (ed) (2016) Handbook of simulation optimization. Springer, Berlin

Gedik N (2016) A new feature extraction method based on multi-resolution representations of mammograms. Appl Soft Comput 44:128–133

Ghosh A, Datta A, Ghosh S (2013) Self-adaptive differential evolution for feature selection in hyperspectral image data. Appl Soft Comput 13:1969–1977

Gysels E, Philippe R, Patrick C (2005) SVM-based recursive feature elimination to compare phase synchronization computed from broadband and narrowband EEG signals in brain–computer interfaces. Signal Process 85(11):2178–2189

Hira ZM, Gillies DF (2015) A review of feature selection and feature extraction methods applied on microarray data. Adv Bioinf. https://doi.org/10.1155/2015/198363

Hong Y, Dong Z (2004) Genetic algorithms with applications in wireless communications. Int J Syst Sci 35(13):751–762

Hossain MS, Muhammad G (2016) Healthcare big data voice pathology assessment framework. IEEE Access 4:7806–7815

Hsu HH, Cheng WH, Ming-Da L (2011) Hybrid feature selection by combining filters and wrappers. Expert Syst Appl 38:8144–8150

Inbarani HH, Azar AT, Jothi G (2014) Supervised hybrid feature selection based on PSO and rough sets for medical diagnosis. Comput Methods Programs Biomed 113(1):175–185

Jain D, Singh V (2018) An efficient hybrid feature selectionmodel for dimensionality reduction. Procedia Comput Sci 132:333–341
Jendoubi T, Strimmer K (2019) A whitening approach to probabilistic canonical correlation analysis for omics data integration. BMC Bioinf 20:1–13
Jiang QY, Li WJ (2015) Scalable graph hashing with feature transformation. In: Proceeding of 24th international joint conference on artificial intelligence, pp 2248–2254
Jinxing C, Yang Y, Li Li BX, Zhang S, Deng C (2017) Maximum relevance minimum common redundancy feature selection for nonlinear data. Inf Sci 409:68–86
Jothi JAA, Rajam VMA (2017) A survey on automated cancer diagnosis from histopathology images. Artif Intell Rev 48:31–81
Kapsoulis D, Tsiakas K, Trompoukis X, Asouti V, Giannakoglou K (2018) Evolutionary multi-objective optimization assisted by matamodels, kernel PCA and multi-criteria decision making techniques with applications in aerodynamics. Appl Soft Comput 64:1–13
Kira K, Rendell LA (1992) The feature selection problem: traditional methods and a new algorithm. Proc AAAI 92(2):129–134
Lee JA, Lendasse A, Verleysen M (2004) Nonlinear projection with curvilinear distances: isomap versus curvilinear distance analysis. Neuro-Comput 57:49–76
Li Y, Ngom A (2013) The non-negative matrix factorization toolbox for biological data mining. Source Code Biol Med 8:10
Li F, Wang J, Chyu MK, Tang B (2015) Weak fault diagnosis of rotating machinery based on feature reduction with supervised orthogonal local fisher discriminant analysis. Neuro-Comput 168:505–519
Lichman M (2013) UCI machine learning repository, University of California, School of Information and Computer Science, Irvine, CA. https://archive.ics.uci.edu/ml/datasets.php
Liu H, Motoda H (2007) Computational methods of feature selection. CRC Press, Boca Raton
Luo Y, Tao D, Ramamohanarao K, Xu C, Wen Y (2015) Tensor canonical correlation analysis for multi-view dimension reduction. IEEE Trans Knowl Data Eng 27:3111–3124
Mahale RA, Chavan SD (2012) A survey: evolutionary and swarm based bio-inspired optimization algorithms. Int J Sci Res 2(12):1–6
Malik ZK, Hussain A, Wu J (2016) An online generalized eigenvalue version of laplacian eigenmaps for visual big data. Neurocomputing 173(2):127–136
Mathias F., Metka B., and Bauer-wersing U. (2018). Navigation system based on slow feature gradients. U.S. Patent Application 15/905,962, filed August 30, 2018
Mazomenos EB, Biswas D, Acharyya A, Chen T, Maharatna K, Rosengarten J, Morgan J, Curzen N (2013) A low-complexity ECG feature extraction algorithm for mobile healthcare applications. IEEE J Biomed Health Inf 2:459–469
McDonnell LA, Remoortere AV, Velde ND, Zeijl RJMV, Deelder AM (2010) Imaging mass spectrometry data reduction: automated feature identification and extraction. J Am Soc Mass Spectrom 21(12):1969–1978
Michaeli T, Wang W, Livescu K (2016) Nonparametric canonical correlation analysis. In: Proceedings of international conference on machine learning, pp 1967–1976
Naji S, Jalab HA, Kareem SA (2019) A survey on skin detection in colored images. Artif Intell Rev 52:1041–1087
Nie F, Huang H, Cai X, Ding CH (2010) Efficient and robust feature selection via joint $\ell 2$, 1-norms minimization. In: Advances in neural information processing systems, pp 1813–1821
Ozdenizci O, Erdogmus D (2019) Information theoretic feature transformation learning for brain interfaces. IEEE Trans Biomed Eng 67:69–78
Patro S, Sahu KK (2015) Normalization: a preprocessing stage. arXiv preprint
Pedram G, Benediktsson JA (2015) Feature selection based on hybridization of genetic algorithm and particle swarm optimization. IEEE Geosci Remote Sens Lett 12:309–313
Raghupati W, Raghupati V (2014) High dimensional data analytics in healthcare. Promise Potential Health Inf Sci Syst 2–3
Ridge K (2005) Kent-Ridge biomedical dataset repository. http://leo.ugr.es/elvira/DBCRepository/index.html
Sacha D, Zhang L, Sedlmair M, Lee JA, Peltonen J, Weiskopf D, North SC, Keim DA (2017) Visual interaction with dimensionality reduction: a structured literature analysis. IEEE Trans Vis Comput Gr 1:241–250
Sorzano C.O, Vargas J, Montano A.P (2014). 'A survey of dimensionality reduction techniques'. preprint arXiv, 1403-2877
Stanojević G, Krivokapić Z (2014) Rare tumors of the colon and rectum in colorectal cancer-surgery, diagnostics and treatment. IntechOpen, Hamilton

Suguna R, Devi MS, Mathew RM (2019) Customer churn predictive analysis by component minimization using machine learning. Int J Innov Technol Explor Eng (IJITEE) 8(8):3229–3233
Sun T, Wang J, Li X, Lu P, Liu F, Luo Y, Gao Q, Zhu H, Guo X (2013) Comparative evaluation of support vector machines for computer aided diagnosis of lung cancer in CT based on a multi-dimensional data set. Comput Methods Programs Biomed 111:519–524
Tan PN (2018) Introduction to data mining. Pearson Education, Chennai
Tao Z, Huiling L, Wenwen W, Xia Y (2018) GA-SVM based feature selection and parameter optimization in hospitalization expense modeling. Appl Soft Comput 75:323–332
Van der Linden C, Dufresne Y (2017) The curse of dimensionality in voting advice applications: reliability and validity in algorithm design. J Elections Public Opin Parties 27(1):9–30
Varghese K, Kolhekar MM, Hande S (2018) Denoising of facial images using non-negative matrix factorization with sparseness constraint. In: Proceedings of 3rd IEEE international conference for convergence in technology (I2CT), pp 1–4
Verónica B, Betanzos A, Amparo M, Sánchez CN (2017) Artificial intelligence, foundations, theory and algorithms feature selection for high-dimensional data. Springer, Berlin
Wang YX, Zhang YJ (2013) Nonnegative matrix factorization: a comprehensive review. IEEE Trans Knowl Data Eng 25(16):1336–1353
Wang J, Tian F, Yu H, Liu CH, Zhan K, Wang X (2018a) Diverse non-negative matrix factorization for multi-view data representation. IEEE Trans Cybern 48:2620–2632
Wang H, Yu D, Li Y, Li Z, Wang G (2018b) Multi-label online streaming feature selection based on spectral granulation and mutual information. In: International joint conference on rough sets. Springer, pp 215–228
Wilms I, Croux C (2015) Sparse canonical correlation analysis from a predictive point of view. Biom J 57:834–851
Xu K, Zhang L, Pérez D, Nguyen PH, Ogilvie-Smith A (2017) Evaluating interactive visualization of multidimensional data projection with feature transformation. Multimodal Technol Interact 1(3):13
Zeren DY, Adhikari N, Wong YK, Aksakalli V, Gumus AT, Abbasi B (2018) SPSA-FSR: simultaneous perturbation stochastic approximation for feature selection and ranking. arXiv preprint
Zeynep A, Thurau C, Bauckhage C (2011) Non-negative matrix factorization in multimodality data for segmentation and label prediction. In: Proceedings of 16th computer vision winter workshop, Austria
Zhang J, Hua H, Wang J (2010) Manifold learning for visualizing and analyzing high-dimensional data. IEEE Intell Syst 25(4):54–61
Zhao C, Gao F (2015) A nested-loop Fisher discriminant analysis algorithm. Chemom Intell Lab Syst 146:396–406
Zhi W, Zhang Y, Chen Z, Yang H, Sun Y, Kang J, Yang Y, Liang X (2016) Application of ReliefF algorithm to selecting feature sets for classification of high resolution remote sensing image. In: Proceedings of 2016 IEEE international geoscience and remote sensing symposium (IGARSS), pp 755–758
Zhou HF, Zhang Y, Zhang YJ, Liu HJ (2019) Feature selection based on conditional mutual information: minimum conditional relevance and minimum conditional redundancy. Appl Intell 49:883–896