Applied Intelligence
https://doi.org/10.1007/s10489-021-02550-9

A comprehensive survey on feature selection in the various fields of machine learning

Pradip Dhal[1] · **Chandrashekhar Azad**[1]

Accepted: 20 May 2021

Abstract

In Machine Learning (ML), Feature Selection (FS) plays a crucial part in reducing data's dimensionality and enhancing any proposed framework's performance. However, in real-world applications, FS work suffers from high dimensionality, computational and storage complexity, noisy or ambiguous nature, high performance, etc. The area of FS is very vast and challenging in its nature. There are lots of work that have been reported on FS over the various area of applications. This paper has discussed FS's framework and the multiple models of FS with detailed descriptions. We have also classified the various FS algorithms with respect to the data, i.e., structured or labeled data and unstructured data for the different applications of ML. We have also discussed what essential features are, the commonly used FS methods, the widely used datasets, and the widely used work done in the various ML fields for the FS task. Here we try to view the multiple comparison experimental results of FS work in different result discussions. This paper draws a descriptive survey on FS with the associated area of real-world problem domains. This paper's main objective is to understand the main idea of FS work and identify the core idea of how FS will be applicable in various problem domains.

Keywords Feature selection · Classification · Machine learning

1 Introduction

In the digital age era, the human's needs for their different kinds of works have become rapidly growing through computer usage. For that, every minute, billion, or even trillions of data is generated. Data has been developed as a valuable asset in recent years, as it's various purposes in different applications. This data now becomes a turn as big data with the property of high dimensionality as nature. So, in this digital age, the big data [1] with high dimensionality features become ubiquitous in various areas like disease analysis or diagnosis, social platform, financial analysis, weather prediction, online educational platform, bioinformatics, and multiple security platforms, etc. The fast development of data raises numerous difficulties for the successful and effective organization of data. We must discover some Artificial Intelligence (AI) and data mining procedures or idea that will automatically gather essential aspects from the collected or stored data. Whenever on these big data, ML or data mining approaches are applied, a severe issue is critically arising, which is a *Curse of Dimensionality (COD)*. The COD [2] is an event that occurs when the high dimensional data doesn't organize, classify, and analyze in a lower-dimensional space; practically, it occurs due to closeness and sparsity of data. Also, in the case of big data, for the massive number of feature spaces, the classification model tends to be overfitted, which may cause the lagging of performance for the unseen data items. High dimensional data [3] significantly increases the computational complexity for the classification model and space complexity for the storage requirements. The primary aspect of classification algorithm or data mining algorithm's performance, to closely related to the important or most valuable features of the dataset. Here the critical element is i) To identify or characterize the relation between the features (or attribute), and ii) The relation or dependency between features and the class (also called outcome or dependent variable). Because not all features are essential, some irrelevant or redundant features may drastically downgrade the classifier's

✉ Pradip Dhal
pradip1780@gmail.com

Chandrashekhar Azad
csazad.ca@nitjsr.ac.in

[1] Department of Compter Applications, National Institute of Technology, Jamshedpur, India

Published online: 23 July 2021

performance. Dimensionality reduction [2] is one kind of process that addresses the issues, which discussed above. FS is divided into two parts i) FS and ii) Feature Extraction (FE) [2]. FS [4] is the process to find out a subset of features from the original set, whereas the FE is the process to extract a new feature subset from the initial subset. Hence the FS and FE techniques increase the classifier's performance, degrades space complexity, time complexity, and storage requirements. In that way, these techniques prove themself efficient dimensionality reduction techniques. In this paper, we only focused on the FS, which is a dimensionality reduction technique. FS's area is very vast, and recently FS task has made a significant impact via performance, time, and space complexity in the various ML regions. In this paper, we have performed a comprehensive survey for the FS task in the different areas of ML. In the following subsection, we have discussed the need for this survey and our contribution to this paper.

1.1 The need for a survey on FS

These are some key aspects due to which we are motivated towards this paper:

1. Dimensionality reduction and its technique, i.e., FS task, are the recent most vital phenomenon in ML algorithms. It is to be needed to identify the core processes within the FS task and analyze the FS framework.
2. To identify the most commonly used features in the FS task, the most widely used FS methods and datasets for the different areas are entirely missing in previous existing surveys.

1.2 Our contributions

Our contributions to this paper are as follows:

1. We have discussed the core processes within the FS task. We also performed a detailed analysis of the FS framework.
2. We have discussed here a detailed design of various FS models. We also classify the different FS algorithms from the data's perspective and discussed the multiple examples.
3. We have presented here the most commonly used features in the various FS task in the different areas of ML. Also, mostly used datasets and FS methods in the various regions of ML.
4. In this paper, we present a complete descriptive survey and classification of various FS tasks in the area of different problems of ML. Based on the FS task's numerous results, we identify the various issues, challenges, and future scope of the various FS methods in the different areas of ML.

1.3 Organization

Section 2 discusses the essential background behind the FS. In Section 3, we have discussed the FS's core processes, the FS framework's analysis, and the various models for the FS task. In Section 4, we have classified the FS algorithms from the perspective of data. In Section 5, we have discussed the most commonly used features in the FS task, the most used FS methods, the most used datasets, and the various kinds of work for FS in the different areas of ML. In Section 6, we have discussed the comparative results of different FS techniques in the various areas of ML. In Section 7, we have discussed the various challenges or issues and future directions of the FS task in different areas of ML. In Section 8, we have concluded the overall work of this paper.

2 The background

In ML and AI, the data becomes a vital aspect of the training and testing phase. The feature or attribute, or property is the most crucial part of the data. The number of features varies for any kind of dataset. In any classification, only the relevant features play a significant role. In the case of classification, the irrelevant features degrade the performance of the overall system. Another critical point is that the classification's performance is not up to the benchmark level for high dimensional data and the larger dataset because of the larger search space. FS is the task of selecting the essential or relevant features so that there will be a chance of improvement in the classification task's performance and accuracy. By eliminating the irrelevant features, the FS task reduces the data's dimensionality, accelerates the classification process, and accuracy or performances increase.

2.1 What are the features or attributes for datasets in ML?

In the area of Pattern Recognition (PR) or ML [2], a feature is an attribute, characteristic or measurable property of an object that is being observed. For example, features can be expressed in different areas of ML such as,

- In Text Classification (TC) system, e.g., to check whether a particular email is spam. The various types of features used in ML algorithms, such as the number of unique URLs in an email, number of words containing letters and numbers, the number of all URLs in an email, the number of words containing only letters, etc.
- In the medical field, to detect skin cancer in the human body, the various types of features can be used in ML

algorithms such as creatinine levels in the blood, neutrophils range, lymphocytes, monocytes, eosinophils, etc.

- For a human body, the fingerprint is the unique pattern of the ridge's shapes and valleys on the finger's surface. Ridge is the single curved section where vally is the area between two adjacent ridges. In the field of image processing, for the fingerprint scanning system, the various types of features can be used in ML algorithms such as the ridge island, ridge ending, ridge dot, ridge enclosure, etc.

2.2 Dimensionality reduction

All the classification problems or tasks are being performed based on the number of features present in their particular area of datasets. All ML algorithm's performance is directly proportional to the number of features in the datasets. Only the relevant features are required for the ML algorithm task. Dimensionality reduction [2] is the process of reducing the number of features from the original set of features so that it could enhance the ML algorithm's performance. The layered structure of dimensionality reduction is shown in Fig. 1.
Question : Why is dimensionality reduction required?
Answer : There are various implications of dimensionality reduction [3] some of them are

- Dimensionality reduction results in the most relevant features so that the accuracy of the ML algorithms increases.
- As the dimension decreases so that the overall ML algorithm's time complexity decreases.
- It also handles the COD phenomena.

The **COD** [2] is the one kind of phenomenon where some specific domain problems like ML, numerical analysis, combinatorics, etc., are solvable only in higher dimensional space rather than the lower-dimensional space.

3 Comprehensive study on feature selection

3.1 Core processes within feature selection

After collecting the dataset, we have no idea, or we are unable to say which features are essential, which are irrelevant, or redundant. In the process of FS, we have to choose the features that result in [3],

- Lesser data compute minimum time for training and testing the ML algorithm process.
- Classifier achieves higher accuracy.

Hence to achieve the above results, there are numerous factors that have to be considered, such as:

1. *How to find the optimal set of features?*
2. *How can we say that these are the optimal set of features, i.e., how to evaluate them?*
3. *In which way new features can be generated, deleting or adding the features from the feature set?*
4. *How to create an FS process as applications independent? i.e., for different types of applications, this process gives the same result.*

The detailed structured framework of the FS process has been described below.

Algorithm 1 Pseudocode for the process of FS.

```
Input: $X$
Output: $f_{optimal}$
Parameters:
$X$=Dataset,
$f_{all}$=All features of the dataset $X$,
$f_{optimal}$= An optimal set of features from the dataset $X$,
$f_{subset}$= The subset of features that is generated via
$SubsetGenerate()$ function through the subset
generate techniques applied ,
$E$=Score value calculated via $SubsetEvaluation()$
function through subset evalution techniques applied,
$T_{condition}$=Termination criteria,
$E_{temp}$, $f_{temp}$=Temporary variable.
$f_{subset} = SubsetGenerate(X, SG_{method})$;
$E = SubsetEvaluation(f_{subset}, X, SE_{method})$;
while $T_{condition}$ do
    $f_{temp} = SubsetGenerate(X, SG_{method})$;
    $E_{temp} = SubsetEvaluation(f_{temp}, X, SE_{method})$;
    if $E_{temp} < E$ then
        $E = E_{temp}$;
        $f_{subset} = f_{temp}$;
    end
end
$f_{optimal} = f_{subset}$
```

```
Function SubsetEvaluation($f$, $X$, $SE_{method}$):
    return $E$;
Function SubsetGenerate($X$, $SG_{method}$):
    return $f_{subset}$;
```

The general framework of the FS task The flow diagram of the detailed structured framework for the FS process is shown in Fig. 2. The pseudo-code of the FS process's algorithmic structure is shown in Algorithm 1 structure. From paper [3], four main steps of the FS process categorized as,

(a) ***Generate feature subset***

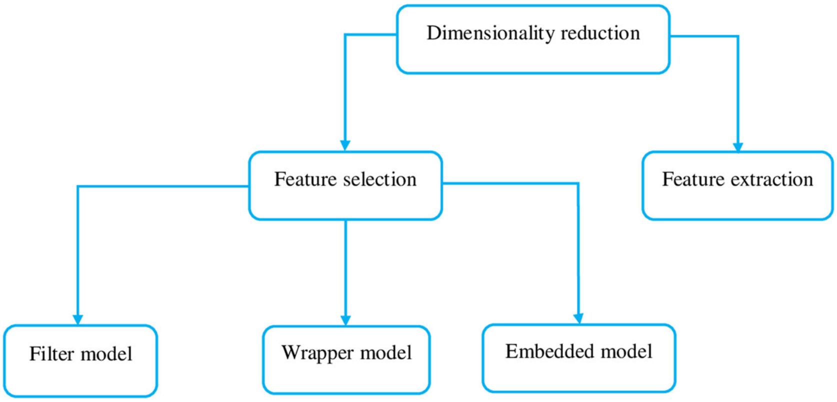

Fig. 1 The layered structure of the FS

(b) ***Feature subset evaluation***
(c) ***Termination condition***
(d) ***Result Validation***

(a). Generate feature subset: For Dataset X of size $(n$ x $f)$, n is the number of rows or instances, and f is the number of columns or features. There are 2^f the possible number of the subset of features is possible. For an f value, small to large, the feature set exponentially increased. Our goal is to find the optimal set of features from the 2^f possible number of subsets. Hence by considering the above questions, the step of feature subset generation [3] can be classified as,

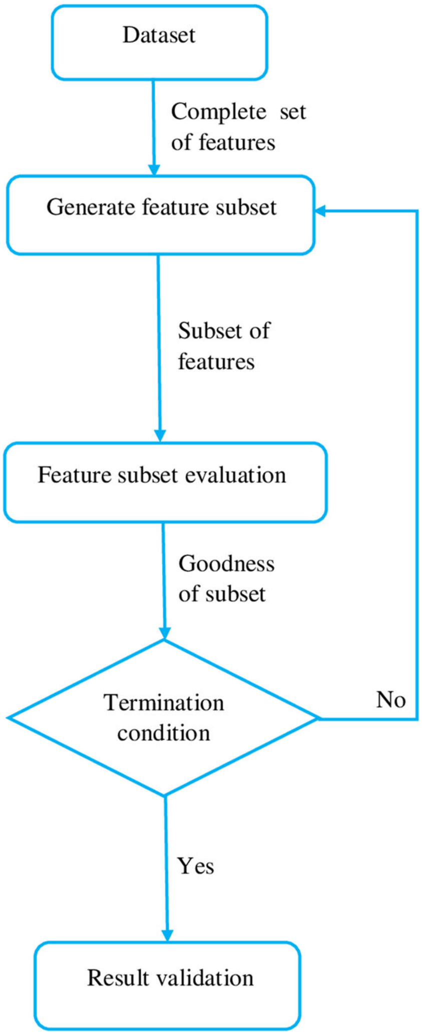

Fig. 2 The general framework for the process of FS

- What is the search methodology we should apply to find the optimal set of features?
- From which direction should we start our search?

Hence in the process of feature subset generation, two following steps are,

- ***Search Methodology (SM)***

  From the above discussion, as a large number of feature subset possible, i.e., $F_{SeachSpace} = 2^f$ where f is a number of features in the dataset, and 2 express two search condition whether to choose the feature subset or not to select the feature subset. In general, as the large subset possible, i.e., 2^f hence the brute force strategies are used for the searching. The most crucial point is that the more you search, the better feature subset you can expect. But from the time complexity point of view, the more you search is very much time-consuming. This is the most challenging part. Hence to achieve the target, i.e., optimal feature subset, and to consider the above part, we can summarize the search [3] as the following categories,

  (i) *Complete Search (CS)* : Complete search will try to find all the possible numbers of feature subsets, and then from that subsets try to fetch the optimal one. Hence the space complexity to fetch the optimal subset is 2^f. However, to find the best solution exhaustive search is needed. But the major disadvantage is that the exhaustive search does not always guarantee the optimal solution. E.g., branch and bound technique found in paper [5] follow this strategy.

  Difference between exhaustive search and CS

Exhaustive search means trying to fetch the possible number of searches but not all searches, but CS implies that all the searches nothing has been left.

(ii) *Heuristic Search (HS)* : This search technique tries to find the optimal solution by enhancing the solution at each step based on some predefined heuristic function or the cost measure. An HS does not always guarantee the optimal solution, but it finds a feasible solution within a considerable amount of space and time complexity, completeness. This search excludes the brute force approach, and for that, it loses the chances to find the optimal solution of the search. This approach runs faster than the other methods that exist. E.g., best-first search, A* algorithm, simulated annealing, and genetic algorithm [6] follow this strategy.

(iii) *Non-deterministic Search (NS)* : In the deterministic algorithm, the output is fixed, i.e., for the searching algorithm, the result is either successful or not. But in the case of the non-deterministic algorithm, the outcome is restricted to a specific set of possibilities, i.e., depends on the conditions in the search result appears. The search follows the randomly selected feature set. Hence unlike the above two search methodology, NS algorithms follow the random strategy to search to achieve the optimal solution. This strategy follows the two most essential criteria,

i) The search will not wait until the search ends.

ii) It does not have the idea that it achieves the optimal set. Instead, it gets a feasible solution if it has.

E.g., the Las Vegas algorithm technique found in paper [7] follow this strategy.

– ***Direction from which SM starts***

No previous idea is given that exactly where the optimal feature subset resides in the search space, so it does not matter that the search will start either from the null set or from the full set of features. The direction from which the feature subset generation starts are related to each other. So the following categories [3] can be stated for the direction of search,

(i) *Sequential Forward Search (SFS)* : Let f_{sfs} be an empty set of features. The SFS started with the empty set f_{sfs}, and at each step, one feature is added in this set. At each step, the most important feature is selected from the unselected set of features based upon some selection criteria. This step is continued until the whole set of features has been covered. Finally, we get a ranked list of features set from high priority to lower priority. The work has been reported in paper [8] follow this direction for FS.

(ii) *Sequential Backward Search (SBS)* : Let f_{sbs} be a full set of features. The SBS started with the complete set f_{sbs} and at each step on feature is removed from the set f_{sbs}. At each stage, the most insignificant feature is removed from the setf_{sbs} based upon some selection criteria. This step is continued until the set f_{sbs} is empty. Finally, we get a ranked list of features set from low priority to high priority. The work has been reported in paper [9] follow this direction for FS.

(iii) *Bidirectional Search (BS)* : This search strategy takes advantage of the above searches, i.e., SFS and SBS. This search starts concurrently by following the SFS and SBS and stops under the following conditions,

1. The two search meets the middle position of the search space.

2. When one or both the search strategy finds the optimal feature set before reach the middle position of the search space.

The work has been reported in paper [3] follow this direction for FS.

(iv) *Random Search (RS)* : In this approach, the optimal feature set is generated randomly. Whether the features will be added or deleted for the optimal subset generation does not follow the fixed way; instead, it follows the random mode. The work has been reported in paper [10] follow this direction for FS.

The comparison between search methodology against the direction of the search for feature subset generation is shown in Table 1.

(b). Feature subset evaluation : After the feature subset generation task is completed, after that the most crucial question is,

What are the different evaluation techniques are there that tell the quality of the optimal feature subsets that are generated from the feature subset generation step?

The evaluation task is very much complicated. It finds the answer to the optimal set features that could increase the accuracy of the classifier. The main aim of the feature subset evaluation task is to increase the classifier accuracy. These are the following measures [3] which are applied in the feature subset evaluation task,

(i) *Divergence measure* : This feature evaluation measure is commonly evaluated the discriminant or divergence capability of the features. It tells how the features

Table 1 Comparison between search methodology *vs* direction of the search concerning Feature subset generation

S. No.	Search methodology	Direction from which search methodology starts			
		SFS	SBS	BS	RS
1	CS	✓	✓	✓	x
2	HS	✓	✓	✓	✓
3	NS	x	x	x	✓

are discriminated against within the classes. A feature with a high divergence distance is more considerable than the low divergence distance or the discriminant power.

Divergence [3] is a special type function that establishes a distance between two probability distribution function. Let us consider for the two class case, $P(X_1|class_1)$ and $P(X_1|class_2)$ be the two probablity distribution function and $D(X_1)$ be the divergence or distance function from $P(X_1|class_1)$ and $P(X_1|class_2)$. Then the feature X_1 will be prefered against X_2 under the following condition holds,

$D(X_1) > D(X_2)$

or similarly we can write as

$D(X_1||X_2) > 0$

One of the important divergence method named as *Kullback–Leibler divergence* [11] is used in fluid mechanics, neuroscience, and machine learning. It quantifies the difference between probability distributions for a given random variable. In mathematically expressed as follows,

$$D_{KL}(a(x)||b(x)) = \sum_{i=1}^{n} a(x_i)\log\left(\frac{a(x_i)}{b(x_i)}\right)$$

Where $a(x)$ and $b(x)$ are the two probability distributions of discrete random variable x and $a(x) > 0$, $b(x) > 0$. $a(x)$ calculates the true or correct observation of data and $b(x)$ calculates theory, model, description, or approximation of $a(x)$. Kullback–Leibler divergence calculates only the distance between two probability distributions, and it is not under the category of the distance measure. It is also not symmetric i.e.

$$D_{KL}(a(x)||b(x)) \neq D_{KL}(b(x)||a(x))$$

The work has been reported in paper [12] follow this measure for feature evaluation.

(ii) *Information Gain (IG) or uncertainty measure* : When the receiver receives all the relevant messages, then IG calculates the amount of uncertainty in the receiver process. If the receiver has an idea of what exactly they are getting, then the amount of the uncertainty level is quite low. But suppose the receiver doesn't know what messages he is getting in that all the messages can get an equal probability. In that case, the amount of uncertainty level is high. Hence in the sense of classification, the messages have been treated as a class. Here IM is the information measure [13] that belongs to the true class, and UT is a function of uncertainty that calculates the messages' uncertainty.

Let us suppose that $P(class_i)$ be the prior class probability where $i = 1, 2, 3, ...n$ and UN is the function of uncertainty. The information gain of the feature x is defined as the difference between the prior uncertainty and the posterior uncertainty of the variable x. Mathematically, it can be written as,

$$IM(x) = \sum_{i=1}^{n} UN\Big(P(class_i)\Big) - E\Big[\sum_{i=1}^{n} UN\Big(P(class_i|x)\Big)\Big]$$

Where,

prior uncertainty = $\sum_{i=1}^{n} UN\Big(P(class_i)\Big)$, and posterior uncertainty=$E\Big[\sum_{i=1}^{n} UN\Big(P(class_i|x)\Big)\Big]$, and the feature of x_1 prefered against x_2 when $IM(x_1) > IM(x_2)$.

One of the most important uncertainty function given by Shannon is,

$$IM(x) = \sum_{i=1}^{n}\left(P\Big(class_i|x\Big) * \log\Big(P(class_i|x)\Big)\right)$$

The work has been reported in paper [14] follow this measure for feature evaluation.

(iii) *Dependency measure* : This measure [3] is trying to quantify the features that, how strongly they are correlated or associated with the class. That's why this dependency measure is also called correlation or association measures. In this feature evaluation process, rather than how the feature value changes through IG by posterior and prior probability, we only concentrate that how the features are strongly associated with the class. If $DM(x_1)$ is dependency measure between the feature x_1 and the class $class$, then the feature x_1 is selected in comparison with x_2 if and only if $DM(x_1) > DM(x_2)$.

One of the important method Hilbert-Schmidt Independence Criterion (HSIC) [15] uses this dependency measure for the feature evaluation process.

This HSIC strategy depends on figuring the Hilbert–Schmidt standard of the cross-covariance administrator of mapped tests to compare Hilbert spaces and is generally used to measure the statistical dependence between random variables. HSIC uses kernels for measuring dependency checking and does not require any density estimation process.

(iv) *Consistency measure* : The above three measures, i.e., divergence, IG, or uncertainty, and dependency measures, try to find the best features concerning the highest feature value. If feature x_1 has the highest feature value than the feature x_2 associated with the class $class_1$ then we will choose x_1. Likewise, these measures finally give us the optimal set. But the main flaw in these processes is that they can't handle the situation when a tie arises. The main contribution in the consistency measure [3] is that it tries to find the minimal number of feature subsets that can classify the problem equivalent to when the full feature set can classify the problem. These measures can eliminate redundant and irrelevant features. In mathematically, we can conclude that,

$$P(class|FullFeatureSet) = P(class|SubsetFeatureSet)$$

One of the best methods is the Las Vegas algorithm [7] uses the consistency measure for the feature evaluation process. It is the probability-based approach, as we know that the HS vulnerable concerning the higher-order correlation here Las Vegas algorithm removes this situation.

(v) *Accuracy measure* : This measure depends on the classifier performance. In FS's whole process, the only goal is to find the optimal feature subset and get the optimal feature subset that only relies upon for the best predictive accuracy. Hence the accuracy measure stands out among all the four previous measures. The work has been reported in paper [16] uses the Support Vector Machines (SVM) classifier to follow this measure for feature evaluation. Likewise, in paper [4] uses Particle Swarm Optimization (PSO) to follows this measure for feature evaluation.

(c). Termination Condition : Termination criteria tell that when the continuation process FS task completes. There are some following conditions for the termination criteria [1],

- There is a threshold point, which may be the maximum number of iterations or some maximum number of features.
- The search process finished.
- Optimal feature subset achieved. (i.e., the classification error rate for the feature subset is less than the allowable error rate for that corresponding framework.)
- For the feature subset generation process, adding or deleting features does not produce a lower error rate of the classifier.

(d). Result Validtion : When the addition or the deletion of features for the feature subset generation process does not produce a lower error rate of the classifier, result validation is the task to confirm that the model's output is acceptable for the real data problem of ML. In other words, we can say that result validates verify that objective has been achieved. For the result validation task following techniques can be applied [3],

- Split Sample Validation
- Cross-Validation
- Bootstrapping Validation

The overall summary of reference papers for each component of the FS framework is shown in Table 2.

3.2 Different models of feature selection

FS task is one kind of data preprocessing task where the redundant and irrelevant features have been omitted to improve the classifier's accuracy and time complexity. First, the input dataset with the set of features is applied to the FS task, and finally, we get the subset of features from this task. That subset of features with the dataset is given to the classifier, and the classifier will calculate the predictive accuracy. Here the irrelevant features are not associated with the class attribute, and the redundant features are features related to the other features. So omitting, irrelevant, and redundant features do not affect the classifier accuracy. By paper [3], the FS task divided into three models,

(i) Filter model
(ii) Wrapper model
(iii) Embedded model

(i). Filter model In this model [2], the FS process takes place by evaluating the quality of the feature subset. The feature subset's quality takes place by some quality measurement techniques and which is entirely independent of any of the classification algorithm. The detailed framework of the filter approach is shown in Fig. 3. The subset of features extracted from the original set of features is taking place by some criteria called the feature relevance measure. After the feature subset extraction, it is input to the classification algorithm; the classifier will calculate the predictive accuracy by using the testing data and the extracted feature subset. In the filter model, the classifier is evaluated at the end of the process rather iteratively more than one. That is why the filter approach is faster than the other two models. Filter model worked on the following two main categories,

Table 2 Techniques behind each component in the FS framework

S.No.	Components	Techniques used		Techniques used in the papers
1	Feature subset generation	Search methodology	CS	[5]
			HS	[6, 17, 18]
			NS	[7]
		Direction from which search methodology starts	SFS	[8, 19, 20]
			SBS	[9, 21]
			BS	[3]
			RS	[10]
2	Feature subset evaluation		Divergence measure	[12, 22, 23]
			IG or uncertainty measure	[13, 14]
			Dependency measures	[24–26]
			Consistency measure	[27–29]
			Accuracy measure	[4, 30]

1. The first category works on measuring each feature's quality (relevance) without considering their interactions with the other features.
2. The second category works on the principle that the aim is to find the subset of features and the features within that subset, the interaction with the other features is maximum.

(ii). Wrapper model The effective way to increase the FS performance is to use the ML algorithm as a performance measure. If our main objective is to reduce the classifier's error rate, then we should go for the ML algorithm as a performance measure. Wrapper model [2] uses ML algorithms to generate the feature subset as a part of the feature evaluation function. Here we will use the subset of features and train the model. The inference that we calculate from the model shows that the features will be added or removed from the subset. This concept is used in the wrapper model. The wrapper model's detailed framework is shown in Fig. 4. In Fig. 4, for the generation of the feature subset, the ML algorithm is used as an evaluation measure. By paper [3], the wrapper model worked via two phases. i) In the first phase, the optimal feature subset generated via the subset search process and the classifier accuracy using the training data. ii) In the second phase of learning and testing, the classifier is tested using the test data through the optimal feature subset.e classifier is tested using the test data through the optimal feature subset.

(iii). Embedded model The filter model is computationally faster than the wrapper model, but it faces the lowest accuracy issue. Instead, a wrapper approach deals with the highest accuracy, but it is computationally slower. Hence, to overcome the filter and wrapper models' problems, we go for the embedded model [3] for the FS task. The embedded model calculates which features are responsible for the accuracy of the model in creating the training model.

The Embedded model performs the FS task between creating the classifier and executing the FS task before creating the classifier. Figure 5 shows the framework for the Embedded model for the FS task. The following kinds of methods are used in the embedded model for FS. The

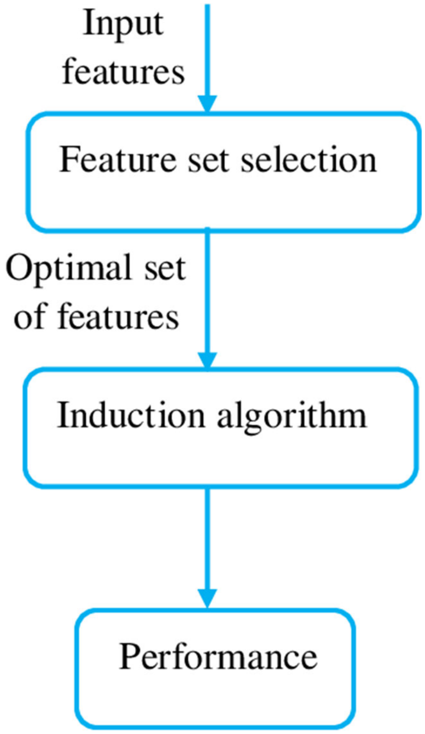

Fig. 3 The general framework for the process of filter model for FS

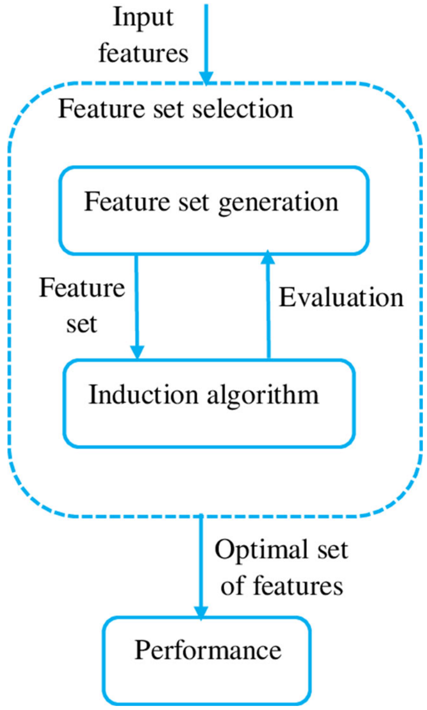

Fig. 4 The general framework for the process of wrapper model for FS

roles of filter, wrapper, and embedded methods in the FS approach are described in Table 3.

Finally, after discussing various FS approaches, the tabular comparison within a filter, wrapper, and embedded models is shown in Table 4.

4 Classification of FS algorithms from the perspective of data

From the perspective of structured or labeled data and unstructured data, the FS algorithms can be categorized as follows,

(i) Statistical measure based FS
(ii) Probability measure based FS
(iii) Similarity measure based FS
(iv) Sparse learning measure based FS
(v) Evolutionary algorithm based FS
(vi) Other methods for FS

(i) **Statistical measure based FS :** Here FS algorithms are based upon various statistical measures. They test function validity using a variety of statistical methods rather than induction algorithms. Furthermore, most statistical-based algorithms examine each feature from the feature set separately. As a result, function redundancy is invariably overlooked during the selection process. The following are some representative statistical measures based on FS algorithms are as follows,

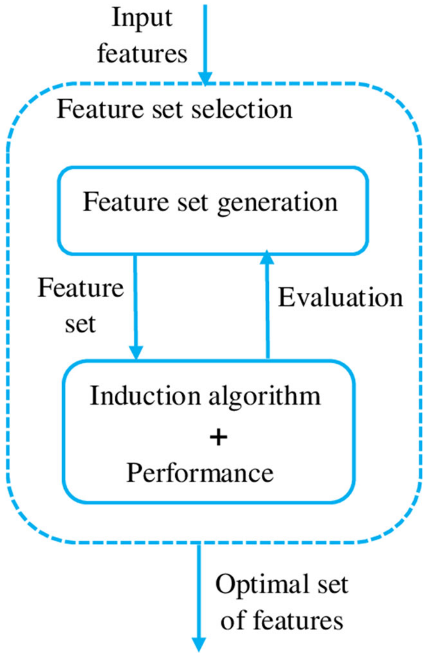

Fig. 5 The general framework for the process of embedded model for FS

(a) ***Chi-Square (χ^2) Statistics*** :

Chi-Square (χ^2) Test is utilized to test the independence of two events. In FS, we will test a specific feature's occurrence, and the occurrence of a specific class is independent. If the two events are dependent, we can use the occurrence of the feature to predict the occurrence of the class. We aim to select the features of which the occurrence is highly dependent on the class's occurrence.

Let us suppose that there are a total n number of instances, and there are two classes, i.e., $True$, $False$. For a feature x, we can use the χ^2 statistics to evaluate its importance in $True$ and $False$. In ML, we can use the χ^2 statistics in the FS technique and ranked the features by calculating the χ^2 scores and then use the top-ranked features for the model training. From paper [2], to calculate the χ^2 score of feature x, we proceed in the following way,

		Class		
		$True$	$False$	Total
Feature	x	a	b	$a+b$
	$\neg x$	c	d	$c+d$
	Total	$a+c$	$b+d$	N

Table 3 Roles of filter, wrapper, and embedded model in FS process

FS method	Role in the FS process
Filter model	– This method goes through various statistical measures for the assignment of rank or score to each feature. – This method selects feature subsets before the learning model. – The process for selecting a feature subset is done only once in the entire building learning model's mechanism. – This method follows the feature relevance property to generate the feature subset. – The process for the generation of feature subset is independent of any ML algorithm. – This method follows the inherent property of the features. Also, it neglects the dependency among the features for the generation of feature subset. – This method doesn't remove the multicollinearity property. It should be removed alternatively by another mechanism.
Wrapper model	– This method goes the accuracy of the model for the assignment of rank or score to each feature. – This method follows the dependency among the features for the generation of the feature subset. – For the generation of feature subset, the classifier runs multiple times. – This method goes for the generation of feature subset by considering the learning model to be applied. – This method selects the most useful features for the generation of the feature subset. – From the inference of the previous learning model, this method decides whether to add or remove the feature from the feature's subset. – This method uses a specific classifier for the evaluation of the quality of the selected features.
Embedded model	– The FS process builds in the training phase. And, this method evaluates the feature subset for the algorithm being trained. – This method uses the usefulness property of the features to generate the feature subset. – This method uses the learning mechanism for the search space. – This method uses the aggregation of the advantages of the filter and wrapper method. – This approach calculates dependency among the features very effectively. Due to the consideration of the classifier, this method selects the relevant features.

a = The quantity of positive occurrences that contain feature x

b = The quantity of negative occurrences that contain feature x

c = The quantity of positive occurrences that does't contain feature x

d = The quantity of negative occurrences that does't contain feature x

$N = (a + b + c + d)$ = Total number of occurrences

$\mathcal{I}$ = Total number of instances

$\mathcal{C}$ = Total number of classes

$$\chi^2 = \sum_{i=1}^{\mathcal{I}} \sum_{j=1}^{\mathcal{C}} \frac{(Obev_{i,j} - Epec_{i,j})}{Epec_{i,j}} \tag{1}$$

Table 4 Comparison within the filter, wrapper, and embedded model

Parameters	Filter approach	Wrapper approach	Embedded approach
Procedure			
Assessment	Statistical test	Cross-validation	Cross-validation
Criteria	Feature subset relevance	Feature subset usefulness	Feature subset usefulness
Search	Through the order of the features (via nested feature subsets or by the individual ranking of features).	Search all possible feature subsets.	The learning process controls the search.
Results	– It is robust against overfitting. – There is a chance to fail, the selection of useful features.	– It is very prone to overfitting. – It selects the most useful features but has massive time complexity.	– Less prone to overfitting. – Comparatively lower time complexity.

Let us suppose that a, b, c, d are the observed values and $Epec_a$, $Epec_b$, $Epec_c$, $Epec_d$ are the expected values respectively. According to the NULL hypothesis that the two events are independent. The expected value of $Epec_a$ can be written as,

$$\frac{Epec_a}{a+c} = \frac{a+b}{N}$$

$$Epec_a = (a+c)\frac{a+b}{N} \tag{2}$$

Likewise we can calculate the values $Epec_b$, $Epec_c$ and $Epec_d$. We can write the value of χ^2 in the following way,

$$\chi^2 = \left(\frac{a-Epec_a}{Epec_a}\right) + \left(\frac{b-Epec_b}{Epec_b}\right) + \left(\frac{c-Epec_c}{Epec_c}\right) + \left(\frac{d-Epec_d}{Epec_d}\right)$$

$$\chi^2 = \frac{(a+b+c+d)(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)}$$

$$\chi^2 = \frac{N(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)} \tag{3}$$

Likewise, we can calculate all the χ^2 values of all the features present in the dataset; after then, we ranked the features' value. Then we select the appropriate features from the top-ranked values. Bahassine et al. [31] used χ^2 for the FS task to classify Arabic text.

(b) ***t - statistics*** :

This is one hypothetical test under the null hypothesis. This test follows the Student's $t-$distribution. It is most ordinarily applied when the test measurement would follow a normal distribution if the estimation of a scaling term in the test measurement were known. It can be utilized, for instance, to decide whether the means for two sets of information are not the same.

Let us suppose that for the two-class problems each instance is classified into either Class 1 or Class 2. By paper [32], for the feature x_i, $t-$statistics can be written as,

$$t - value(x_i) = \frac{|(mean_1 - mean_2)|}{\sqrt{\left(\frac{var^2_{i,1}}{num_1}\right) + \left(\frac{var^2_{i,2}}{num_2}\right)}} \tag{4}$$

Where,

num_1 = Total number of instances for the class 1

num_2 = Total number of instances for the class 2

$mean_1$ = Mean value of the feature x_i for the class 1

$mean_2$ = Mean value of the feature x_i for the class 2

$var_{i,1}$ = Variance value of the feature x_i for the class 1

$var_{i,2}$ = Variance value of the feature x_i for the class 2

Similarly, we can calculate the t value for all the features; after then, we ranked the feature values. Then we select the appropriate features from the top-ranked values. Ravisankar et al. [33] used a t-test for the FS task to fraud detection in the financial statement.

(c) ***Analysis of variance (ANOVA)*** :

Ronald Fisher develops it, also called Fisher's analysis of variance. It is the extension of the $t-$ and $z-$tests. ANOVA is used to analyze the experiment of comparison for those in which only the differences are of interest. The ratio of two variances can obtain statistical significance. It is independent of the experimental observations' distinct alterations; adding or multiplying any constant doesn't change the significance. Hence the ANOVA result is independent of the scaling error or the constant bais. ANOVA is a statistical test that analyses the variations between and among the group means in a particular sample. It is used to evaluate whether the means of one or more population are equal or not, and for that, it follows the $t-$test for the means of population.

The F-test is utilized for comparing the result of the factor of total deviation. In a one way ANOVA, statistical significance is tested via Comparing the $F-$test statistic, which is written as via paper [2],

$$F = \frac{MSSC}{MSSE} \tag{5}$$

Where,

F = Coefficient of ANOVA

$MSSC$ = Mean sum of squares between the classes

$MSSE$ = Mean sum of squares within the classes

$\mathcal{C}$ = Total number of classes

n_i = The total number of samples for the class i

N = The total number of sample points = $\sum_{i=1}^{\mathcal{C}} n_i$

$x_{j,i}$ = This is the j^{th} sample point for the class i

X_i = The sum of all the numbers for the class i = $\sum_{j=1}^{n_i} x_{j,i}$

$\overline{X_i}$ = The sample mean for the class i = $\frac{X_i}{n_i}$

$\overline{X}$ = The mean of of the sample points = $\frac{\sum_{i=1}^{\mathcal{C}} X_i}{N}$

var_i^2 = The sample variance for the class i = $\frac{\sum_{j=1}^{n_i}(x_{j,i}-\overline{X_i})^2}{(n_i-1)}$

$MSSC = \frac{\sum_{i=1}^{\mathcal{C}}(\overline{X_i}-\overline{X})^2}{\mathcal{C}-1}$, $MSSE = \frac{\sum_{i=1}^{\mathcal{C}}(n_i-1)var_i^2}{N-\mathcal{C}}$

In the paper [34], FS has been performed via $p-$value based on the $F-$test ANOVA technique.

(ii) **Probability measure based FS :** Here FS algorithms are based upon various probability measures. To minimize feature redundancy and maximize feature relevance, several probability measures parameters are being proposed. Furthermore, the bulk of probability principles can only be generalized to discrete variables. As a consequence, this family of FS algorithms can only work with discrete data. Some data discretization techniques are needed before using continuous feature values. The following are some representative probability measures based on FS algorithms are as follows,

(a) ***Mutual Information (MI)*** :

The measurement of the uncertainty of the discrete random variable is called entropy. By paper [2], the entropy for discrete random variable U can be written as,

$$H(U) = -\sum_{u_i \in U} Pr(u_i)log(Pr(u_i))$$

Where u_i represents the specific value for the random variable U. $Pr(u_i)$ is the probability of u_i for each possible event $u_i \in U$(all possible events). The conditional entropy of U given that the discrete random variable V is written as,

$$H(U|V) = -\sum_{v_j \in V} Pr(v_j) \sum_{u_i \in U} Pr(u_i|v_j)log(Pr(u_i|v_j))$$

$Pr(v_j)$ denotes the prior probability of v_j and the $Pr(u_i|v_j)$ denotes the conditional probability of u_i given that v_j.

MI is the process to estimate the mutual dependencies between the two discrete random variables. The MI within U and V are used to calculate the amount of information exchange via U and V together,

$$MI(U;V) = H(U) - H(U|V) = \sum_{u_i \in U} \sum_{v_j \in V} Pr(u_i, v_j)log\left(\frac{Pr(u_i,v_j)}{Pr(u_i)Pr(v_j)}\right) \tag{6}$$

Where, $Pr(u_i, v_j)$ is the joint probability of u_i and v_j.

Let f_s be the set of features on f_{all} which is the set of all features, and C be the class labels. Then the amount of information gathered by the feature $f_i \in f_{all}$ for class C must have the maximum information compared to all the features in the subset f_s, and then we can select the feature f_i. Tang et al. [35] used MI for the FS task to classify the textual document.

(b) ***Information Gain (IG)*** :

IG calculates the importance of the feature in the corresponding feature vector. IG is an entropy-based FS method, and it is broadly used in various filed of ML applications. IG can be determined by determining how much of a feature or attribute can classify the information for the measurement of the importance of the lexical items in the classification.

(c) ***Gini Index (GI)*** :

Let X be the dataset and $X = (x_1, x_2, x_3, ..., x_n)$ be the set of n features and x_i feature have m different feature values. Then $\mathcal{Y}$ be the set of instances that values will be smaller and equal to the feature value x_i, and $\overline{\mathcal{Y}}$ be the set of instances that values will be larger than the value x_i. By paper [2], the score of GI for the feature x_i can be written as,

$$G(x_i) = Pr(\mathcal{Y})\left(1-\sum_{i=1}^{\mathcal{C}} Pr(C_i|\mathcal{Y})^2\right)+ Pr(\overline{\mathcal{Y}})\left(1-\sum_{i=1}^{\mathcal{C}} Pr(C_i|\overline{\mathcal{Y}})^2\right) \tag{7}$$

Where $Pr(\mathcal{Y})$ is the probability of the set of instance $\mathcal{Y}$. and $Pr(\overline{\mathcal{Y}}) = 1 - Pr(\mathcal{Y})$. The conditional probability $Pr(C_i|\mathcal{Y})$ denotes that the probability of class C_i has given $\mathcal{Y}$. Azam et al. [36] used GI for the FS task to classify the textual document.

(d) ***minimum Redundancy–Maximum Relevance (mRMR)*** :

By paper [2], this method selects only that particular features that can have the minimum redundancy, which means the features are maximum nonsimilar to each other and highly relevant to the target class. Both the points, i.e., mRMR, are entirely dependent upon the MI. In this approach, at first, the MI between the target variable and the candidate variable has been evaluated, i.e., the relevant term. After that, the average MI is calculated between the candidate variable and the selected variables, i.e., redundant term. The mRMR value for the feature is calculated, which is the difference between the relevant term and redundant term. More the mRMR value for a feature most important feature. Let, f_{all} be the total number of feature subsets. Currently, we have an f_n number of feature subsets available, where n is the number of features in the f_n feature subset. Then the i^{th} feature is chosen from the subset $(f_{all} - f_n)$. Then according to mRMR, we can write as,

$$\underset{x_i \in (f_{all} - f_n)}{Max} \left[MI(x_i; t) - \frac{1}{n} \left(\sum_{x_k \in f_n} MI(x_i; x_k) \right) \right]$$

Where MI is the mutual information value. Togaçar et al. [37] used mRMR for the FS task to detect lung cancer from chest CT images.

(e) ***Correlation-based Feature Selection (CBFS)*** :

The CBFS process evaluates a subset of features based on the evaluation function. The motto of that evaluation function is that the subset of features maximal correlated with the class and yet non-correlated. Must be ignored redundant and irrelevant features because they have a lower correlation with the class. By paper [2], the score of merit for the subset of features f_n that consist of n features can be written as,

$$MeritScore_{f_n} = \frac{n\overline{T_{cx}}}{\sqrt{n + n(n-1)\overline{T_{xx}}}} \tag{8}$$

Where, $\overline{T_{cx}}$ is the average value of all the features and class correlation, and $\overline{T_{xx}}$ is the average value of all the features and features correlation. The symmetric uncertainty SU is one kind of measure is used to calculate the potential of correlations among the features and the strength of the predictive nature of features for the target class.

$$SU(A; B) = 2\left[\frac{MI(A; B)}{H(A) + H(B)}\right] \tag{9}$$

Here, SU has been normalized in the range of 0 and 1, and it calculates the correlation among features and target class. After that, the features are weighted according to SU value. The maximum value of SU, the better quality of the feature. Shao et al. [38] used CBFS for the FS task to classify multi-label subcellular bio-images.

(iii) **Similarity measure based FS :** Here FS algorithms are based upon various similarity measures. To determine the importance of features, different FS algorithms use various types of parameters. A family of methods for determining feature significance based on their ability to maintain data similarity. They belong to the similarity measures based approaches. The following are some representative similarity measures based on FS algorithms are as follows,

(a) ***Fisher Score (F-Score)*** :

This technique selects only particular features, such that within the same class, the feature values are similar, and in the nonsimilar classes, the feature values are completely different. By paper [39], F-Score for the feature x_i can be written as,

$$F - Score(x_i) = \frac{\sum_{k=1}^{\mathcal{C}} num_k (mean_{i,k} - mean_i)^2}{\sum_{k=1}^{\mathcal{C}} num_k * var_{i,k}^2} \tag{10}$$

Where,

$\mathcal{C}$ = Total number of classes

num_k=Total number of instances for the class k.

$mean_i$=Mean value of the feature x_i.

$mean_{i,k}$=Mean value of the feature x_i for the class k.

$var_{i,k}$=Variance value of the feature x_i for the class k.

Similarly, we can calculate the $F - Score$ value for all the features; after then, we ranked the feature values. Then we select the appropriate features from the top-ranked values. Liu et al. [40] used $F - Score$ for the FS task to classify emotions from the speech signal.

(b) ***Relief*** :

From paper [2], the Relief algorithm has been developed by Kira and Rendell, which is based on instance-based learning. In the FS framework, the feature subset evaluation part, the Relief algorithm calculates each feature 'relevance' or 'quality' to the target concept. This algorithm calculates the feature weights, which are $\mathcal{W}[x_i]$=weight of the feature x_i or the feature score, which varies from -1(worst) to 1(best). At first, the Relief algorithm was only used for binary classification problems, and there is no scope to handle the missing values. These days the Relief algorithm has been extended for the multiclass problems and the continuous point values. The pseudo-code of the original Relief algorithm is described in Algorithm 2. In the Relief algorithm stated in Algorithm 2, it first chooses a n random number of samples from the original dataset. Then from that, i.e., n it chooses one random sample r_i, finds its nearest neighbor from the same class, i.e., H, and finds its nearest neighbor from the different class, i.e., M. It updates the weight vector i.e.

$\mathcal{W}[x_i]$ for the feature x_i that depends upon the the sample r_i for class H and Class M. The function $\Phi(x_i, d_1, d_2)$ is stated as follows, It calculates the difference of the feature x_i between class d_1 and class d_2.

For the continuous set of features(numerical or ordinal values) the Φ can be written as,

$$\Phi(x_i, d_1, , d_2) = \frac{|Val(x_i, d_1) - Val(x_i, d_2)|}{max(x_i) - min(x_i)}$$

For the discrete set of features(nominal or categorical values), the Phi can be written as,

$$\Phi(x_i, d_1, , d_2) = \begin{cases} 0 & Val(x_i, d_1) = Val(x_i, d_2) \\ 1 & otherwise \end{cases}$$

Lu et al. [41] used Relief for the FS task to find breast cancer.

Algorithm 2 Pseudocode for the Relief algorithm.

```
Data: Given the dataset $X$
Result: The feature vector $\mathcal{W}$ contain the fitness scores
        of all the features
$f_{all}$=Total number of features;
$N$=Total number of training samples;
$n$=The number of total random training samples which
    is taken from $N$ training samples;
for each feature $i = 1$ to $N$ do
    Initialize weight of the feature ;
    $\mathcal{W}[x_i]$=0
for $i = 1$ to $n$ do
    Select randomly a target sample $r_i$;
    find a nearest hit $H$ and nearest miss $M$ (samples);
    for $j = 1$ to $f_{all}$ do
        $\mathcal{W}[x_j] = \mathcal{W}[x_j] - \frac{\Phi(x_j, r_i, H)}{n} + \frac{\Phi(x_j, r_i, M)}{n}$
```

(c) **Sparse learning measure based FS :** Here FS algorithms are based upon various sparse learning measures. Sparse-learning-based approaches strive to reduce fitting errors while also incorporating some sparse regularisation terms. Many feature coefficients are forced to be small or zero by the sparse regularizer, and the corresponding features can then be easily removed. Due to their high performance and interpretability, sparse-learning-based methods have gained a lot of attention in recent years. The following are some representative sparse learning measure based on FS algorithms are as follows,

(iv) ***Least Absolute Shrinkage and Selection Operator (LASSO)***

Algorithm 3 Pseudocode for GA based FS.

```
Input: $X$
Output: $f_{optimal}$
Parameters:
$Z$=Total population of dataset $X$,
$f_{optimal}$=Optimal set of features from the dataset $X$,
P=Individual of the population,
$MaxIter$=Termination criteria.
$i \leftarrow 0$;
Initialize $P_i$ as random individuals from the population
    $Z$;
$CalculateFitness(Z, P_i)$;
while $i < MaxIter$ do
    Select the individuals from $P_i$;
    Crossover the individuals;
    Mutate the individuals;
    $CalculateFitness(Z,$modified individuals$)$;
    $P_{i+1} \leftarrow$ select the individuals through Roulette
        wheel selection;
    $i \leftarrow i + 1$;
end
$f_{optimal}$ = Features with the maximum fitness value
```

LASSO is a regression analysis technique used in the regularization and selection of the variable to extend the prediction model's accuracy and interpretability. LASSO is a regression for the linear model, which is achieved by adding a penalty against complexity to overcome the model's overfitting or variance by adding more bias. Zini et al. [42] used LASSO for the FS task in the facial recognition system.

- **Evolutionary algorithm based FS :** Here FS algorithms are based upon various evolutionary algorithms. Evolutionary computation (EC) is a subfield of AI and soft computing that explores a family of global optimization algorithms inspired by biological evolution. EC techniques are standard in computer science because they can deliver highly optimized solutions in various problem settings. The following are some representative evolutionary algorithms that are used mainly in the FS task are as follows,

  (a) ***Genetic Algorithm (GA)*** : John Holland has introduced GA [6] in 1960, based on the Darwin theory of evolution. GA is inspired by the genetics and natural selection that belongs to the evolutionary algorithms. It is generally used to give the optimal solution to search problems and optimization problems. GA mimics natural selection procedure; it means that only those categories of species will be eligible to go to the next generation that can adopt changes in their environment, and they can survive and replicate. It means that it mimics the "fittest candidate" from the individual of the successive generation to solve the problem.

Here the word generation consists of a group or population of individuals where each individual is a member of search space and possibly the solution to the problem. The FS with correspondent to the GA is shown in Algorithm 3. Some essential terms correspondent to the GA discussed below,

(i) ***Crossover*** : In GA, the crossover is also called recombination, is a genetic operator that combinate the two genetic chromosomes, e.g., parents, to produce the new offspring.

(ii) ***Mutation*** : In GA, the mutation is a genetic operator used to maintain and introduce diversity from the generation of a population of genetic chromosomes to the next generation. It just altered from its initial state, one or more gene values in a chromosome to get a new solution.

(iii) ***Population*** : The population is the collection of chromosomes. In GA, the population is a subset of solutions for the current generation.

(iv) ***Fitness function*** : GA must be associated with the fitness function; it calculates the score or the individuals' rank in the population.

(v) ***Selection*** : In GA, the selection of the individuals is performed by the Roulette wheel function. In roulette wheel, the probability of the individual $'a'$ is selected can be written as,

$$P(X = a) = \frac{fit(a)}{\sum_{k=1}^{n} fit(k)} \tag{11}$$

Where $fit(a)$ is the fitness score of $'a'$.

Vijayanand et al. [18] used GA for the FS task to Intrusion Detection System (IDS) in the wireless mesh network.

Algorithm 4 Pseudocode for the PSO based FS.

```
Data: Dataset $X$
Result: Fitness values of all the features
Parameters:
$n$=The total number of features of dataset $X$.
for each particle $i = 1$ to $n$ do
    Initialize particles;
maximum number of iterations for each particle $i = 1$ to $n$ do
    Calculate the fitness value ;
    if If the fitness value is maximum than the previous $Pbest^{local}$ value then
        Set current fitness value as the new $Pbest^{local}$ value;
Choose the particle with the max fitness value according to $Pbest^{global}$ value;
for each particle $i = 1$ to $n$ do
    Calculate the particle velocity via (12);
    Calculate the particle position via (12);
```

(b) ***PSO*** :

By paper [4], PSO is one kind of optimization method inspired by the bird flocking's social behavior. Let us consider the following situation: a group of birds is searching randomly for food in an area. Only a small amount of food resides in a particular location of that area. All the birds don't have any idea what is the exact location of food in that area. But they have an idea of how far the food is in each iteration. Hence, the best strategy to locate the food is to follow the bird nearest the food.

In the aspect of PSO, every solution is a $bird$ in the search space. Let us suppose that it is $particle$. In PSO, all the $particles$ must have a fitness value, which is evaluated by a fitness function that must be optimized, and the particle must have $velocity$ for flying the $particle$. In FS, we apply the PSO in the following way; first, the PSO initialized with the particles' random solutions. After then, the particles searched for the optimum solution by updating the generations. Each particle have the following three values,

vel=Velocity of the particle.

pos=Position of the particle.

$Pbest^{local}$= The best fitness value of the particle.

And the $Pbest^{global}$ is the value obtained by particles.

i.e. $Pbest^{global}$=The best fitness value of any of the particles.

In dataset X for the n dimensional space, where the features are represented as x_1, x_2, $x_3, ...x_n$. Let us suppose in the contrast of PSO each features are represented $particles$. Hence the position vector for x_i is represented as $(pos_{i,1},\ pos_{i,2},\ pos_{i,3}, ..., pos_{i,n})$ and the velocity vector is represented as $(vel_{i,1},\ vel_{i,2},\ vel_{i,3}, ..., vel_{i,n})$. The following two equations have been used to calculate the position and velocity of each particle.

$$vel^{new}_{i,n} = vel^{old}_{i,n} + c1 * rand_1 * (Pbest^{local}_{i,n} - pos^{old}_{i,n}) + c2 * rand_2 * (Pbest^{global}_{n}) - pos^{old}_{i,n}) \tag{12}$$

$$pos^{new}_{i,n} = pos^{old}_{i,n} + vel^{new}_{i,n} \tag{13}$$

After getting all the fitness values we ranked the feature values. Then we select the appropriate features from the top-ranked values.

The pseudocode for the PSO based FS work is shown in Algorithm 4. Kushwaha et al. [43] used PSO for the FS task to classify the textual document.

(c) ***Bacterial Colony Optimization (BCO)*** : The life-cycle model of BCO [44] employs an internal circulation mode, which contributes to the search's high computational cost. The BCO uses predefined steps to carry out previous "foraging search strategies." In general, BCO uses the rule condition to limit the number of operations associated with reduction, replication, and dispersal circling in the chemotaxis phase. In BCO, bacteria use various communication topology frameworks to understand, including dynamic neighbor-oriented study (or randomly oriented study) and community-oriented study. The chemotaxis procedure is made up of two steps: running and tumbling, which can be written as:

Process of tumbling : In tumbling, a stochastic direction participates in the actual swimming phase. As a result, the search orientation in tumbling is influenced by both chaotic and optimum searching directors, and the positions of each bacterium are modified as a result.

$$Pos_i(t) = Pos_i(t-1) + c_i\left[k_i\left(g_{best} - Pos_i(t-1)\right) + (1-k_i)\left(p_{best_i} - Pos_i(t-1)\right) + Tur_i\right] \tag{14}$$

Process of running : No turbulent director is assisting in the running phase to influence the bacteria running in the optimal direction.

$$Pos_i(t) = Pos_i(t-1) + c_i\left[k_i\left(g_{best} - Pos_i(t-1)\right) + (1-k_i)\left(p_{best_i} - Pos_i(t-1)\right)\right] \tag{15}$$

Where, $k_i \epsilon [0, 1]$, Tur_i is the i^{th} bacterium's turbulent path variance, p_{best_i} is the personal best position and g_{best} is the global best position of the i^{th} bacterium. And, c_i is the step size of chemotaxis which is written as,

$$c_i = c_{min} + \left(\frac{I_{max} - I_j}{I_{max}}\right)^n * (c_{max} - c_{min}) \tag{16}$$

Where, I_{max} is the maximum number of iteration, I_j is the current iteration. And, c_{max}, c_{min} is the maximum and minimum stepsize of chemotaxis. When $n = 1$, the chemotaxis move strategy is linearly decreasing. Aside from that, the size of the chemotaxis is changing in a nonlinear decreasing approach. Wang et al. [45] propose a BCO based FS approach that uses in complex structures fault diagnosis.

(d) ***Ant Colony Optimization (ACO)*** :

$Biological - idea$: Ants live in a large group, which is called colonies. An ant can find the shortest path from its colony to the source of the food through pheromone trails. ACO [46] is a population-based search concept applied for the optimization problem whose idea is based upon the biological idea of ants. *How the ants perform this?*

- They don't use vision for this, or they almost blind.
- Some real ants find the shortest path between the colony to the food source.
- The ants decay pheromone trails, which is some chemical left on the ground by them, which acts as a signal for the other ants.
- If an ant decides to follow the pheromones, it also left some pheromones and then followed it.
- More ants follow pheromone than more potent pheromones, and thus it likely more ants follow it.
- The pheromones have the property it destroys within few minutes.
- Hence the pheromone trail builds on the smallest path faster because it does not have

time to be destroyed, and the ants follow this path.

The idea for artificial ACO

- Must have some data structure.
- It can sense the environment like pheromone trails.
- Apply some discrete-time like to destroy pheromone trails.
- Can applicable for optimization problems.

Steps for ACO algorithm:

Algorithm 5 Pseudocode for the process of SVM-RFE.

```
Input: $X$
Output: $f_{optimal}$
Parameters:
$X$=Dataset,
$f_{optimal}$=Optimal set of features from the dataset $X$,
$n$=Total number of features,
$F$=Set of all features,
$Y$=$y_1, y_2, ..., y_n$ where $y_1, y_2, .., y_n$ are the instances of the dataset $X$,
$C$=$c_1, c_2, ..., c_n$ where $c_1, c_2, .., c_n$ are the class labels,
$R$= Empty list for the rank value of each features,
$SVMTrain()$ function optimize the weight vector for the features concerning the weight function.
while $F$ do
    $\beta$=$SVMTrain(Y, C)$ ;
    $W$=$\sum_{i=1}^{n} \beta_i * c_i * y_i$= Calculate the weight vector;
    for each feature $i = 1$ to $n$ do
        $e_i = (w_i)^2$=Compute the rank criterion
    end
    $f$=$min(e)$=Find the feature with smallest rank;
    $R$=Update the rank list features by adding rank of feature $f$;
    $F^1$=eliminate the feature with the lowest ranking from $F$;
    $F \leftarrow F^1$;
end
$f_{optimal}$ = Features with the top ranked list from $R$
```

Assume that ACO is an optimization problem of the Travelling Salesman Problem (TSP).

i. Initialize the positions of each ant and also initialize each pheromone value of each edge of the tour.
ii. There must be a table that holds the travel history for each ant, where the table's first entry must be the starting town from where it starts traveling. Assume an ant moved from one town to another with some probability $p(a, b)$ and left some pheromone trails weights. where, $p(a, b)$ is the probability for the edge ab from a to b.
iii. All ants complete their n moves, i.e., total moves, and fill up their travel history. Then it calculates its L_t and $\delta^t_{a,b}$. After that, extract the shortest path and nullify the table entry for travel history. where, where L_t is the total distance for t town tour and $\delta^t_{a,b}$ is the pheromone trail laid by the s ants on the edge ab for t town tour.
iv. This process continues until all the ants follow the same path or the maximum threshold for the pheromone trail's decay.

Kanan et al. [46] used ACO for the FS task in the facial recognition system.

1. **Other methods for FS :** Hybrid FS methods aim to create a group of feature subsets from various FS algorithms and aggregate the results. Hybrid FS methods have two steps: (1) construct a set of different FS results and (2) aggregate different outputs into a consensus result. An example of the hybrid FS algorithm is,

(vi) ***SVM - Recursive Feature Elimination (SVM-RFE)*** : In classification algorithms, SVM is one of the best classifiers among others due to its computational power. Instead, SVM can handle non-linear decision boundaries. We will calculate each feature value in the FS task, either each iteration, added, or features eliminated at each iteration to compute the feature subset. But there is a significant issue when the feature elimination method has removed several features to calculate the feature subset. This condition can be handled by the RFE technique, which is given below,

- Calculate the ranking procedure for all features.
- Train the classifier by gaining the optimization of weights of the features concerning the cost function.
- Eliminate the features with the lowest ranking criteria.

This is an example of the backword feature elimination process. In the computational complexity case, it is the most appropriate method rather than the several features that are eliminated at a time.

The SVM-RFE algorithm is described in Algorithm 5. Li et al. [19] used SVM-RFE for the FS task to identify tumors from endoscopy images.

Deep learning methods are now widely used and effective in a variety of real-world applications. Although deep learning is mostly used for feature learning, several attempts to use deep learning FS techniques have been made.

5 FS on various areas of ML

After going through the various research papers via different publications, we divided most of the FS work into two categories: structured or labeled data and unstructured data. Fig. 6 shows a clear idea about most of the FS works in various areas of ML via our survey. Here we have discussed the FS task in the three perspectives for the structured or labeled data and the unstructured data. These perspectives are:

- **Question 1:** What are the common types of features for the FS task in this area?
- **Question 2:** What are the most common types of datasets used for the FS task in this area?
- **Question 3:** What are the most common types of methods that researchers usually used along with their advantages?

We have tried to answer these four perspectives in the following subsections (Fig. 6).

5.1 FS on the structured or labeled data

Labeled data is a designation for pieces of data labeled with one or more labels marking specific properties or attributes or classifications or artifacts found within. Labels make this knowledge directly useful in some forms of ML, such as ML configurations that are supervised. Table 5 shows the most common types of datasets used for the FS task for the structured or labeled data.

Table 6 shows the most common types of features used for the FS task for the structured data. Table 7 shows some of the FS tasks for the structured or labeled data.

Table 8 shows the most commonly FS methods used for the structured or labeled data with their advantages.

5.2 FS on the unstructured data

Unstructured data is data that is not ordered, doesn't have the predefined class label. Example: logs for Word, PDF, email, image, media, etc. Analysis of unstructured data is a crucial part of the ML process. From Fig. 6 in the case of the unstructured data, we will discuss the FS work in the following perspective:

1. From the perspective of Natural language processing (NLP)
2. From the perspective of Signal processing
   (a) Image processing
   (b) Speech processing

1. From the perspective of NLP In NLP, most of the research done for the FS work for text classification, sentiment analysis tasks. Table 9 shows the most common types of datasets used for the FS task in the area of NLP. Table 10 shows the most common types of features used for the FS task in the area of NLP.

Table 11 shows some of the FS tasks in the area of NLP.

In Table 12, we have shown the most commonly FS methods used in the area of NLP with their advantages.

2. From the perspective of Signal processing:

(a) Image processing: In image processing, most of the research done for the FS work for biometrics, disease analysis, image classification tasks. In Table 13, we have shown the most common types of datasets that are used for the FS task in the area of image processing. Table 14 shows the most common types of features used for the FS task in the area of image processing.

Table 15 shows some of the FS tasks in the area of image processing.

In Table 16, we have shown the most commonly FS methods used in the area of image processing with their advantages.

(b) Speech processing: In speech processing, most of the research done for the FS work for emotion recognition, disease analysis tasks. In Table 17, we have shown the most common types of datasets that are used for the FS task in the area of speech processing. In Table 18, we have shown the most common types of features that are used for the FS task in the area of speech processing.

Table 19 shows some of the FS tasks in the area of speech processing.

In Table 20, we have shown the most commonly FS methods used in the area of speech processing with their advantages.

6 Results and discussion

We have discussed various FS techniques to the corresponded area of applications. The result of different FS methods depends on different datasets and also different

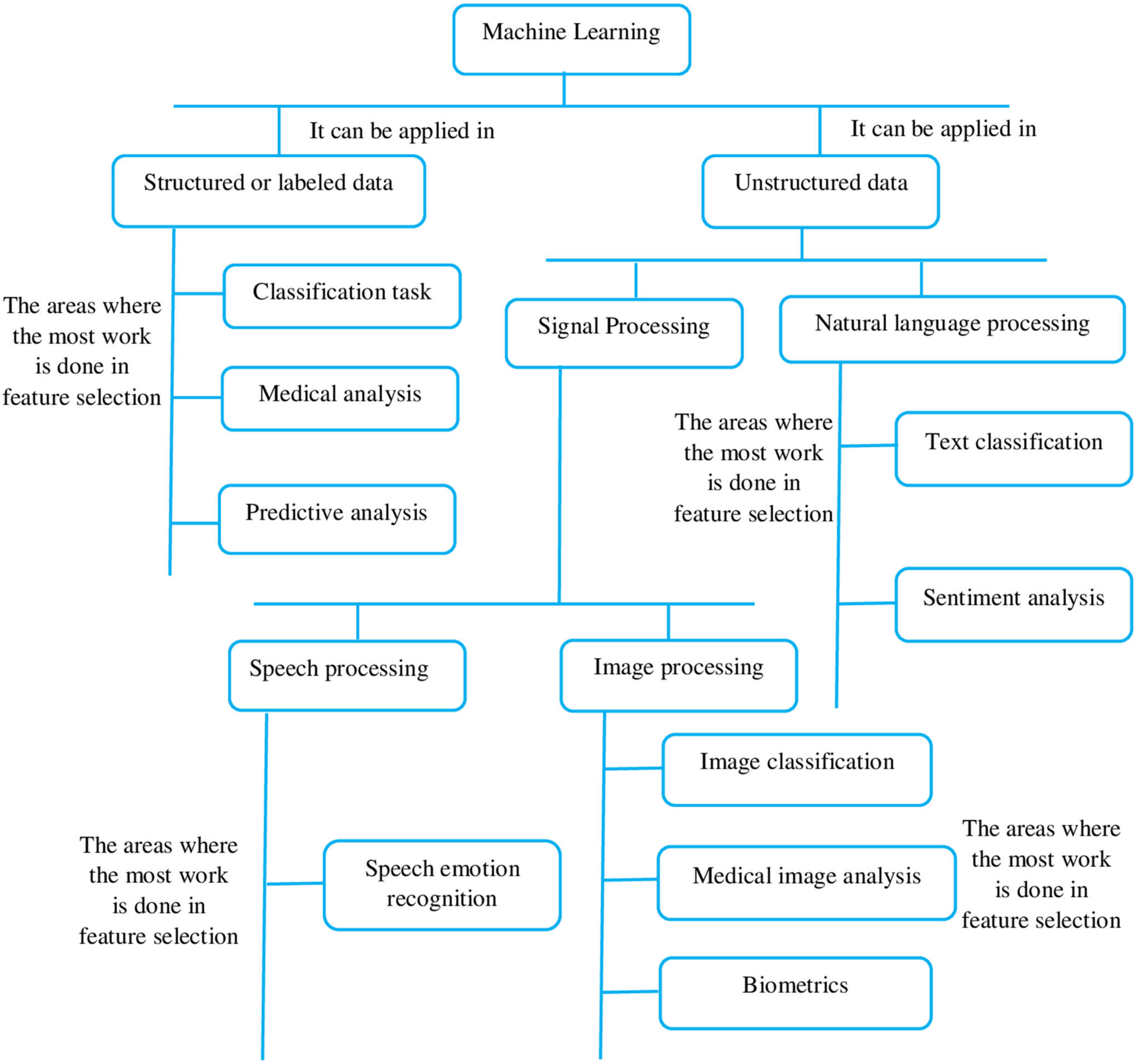
Machine Learning
It can be applied in
It can be applied in
Structured or labeled data
Unstructured data
The areas where the most work is done in feature selection
Classification task
Medical analysis
Predictive analysis
Signal Processing
Natural language processing
The areas where the most work is done in feature selection
Text classification
Sentiment analysis
Speech processing
Image processing
Image classification
The areas where the most work is done in feature selection
Speech emotion recognition
Medical image analysis
The areas where the most work is done in feature selection
Biometrics

Fig. 6 According to our survey, graphical representation of most research in the FS work in the various areas of ML

Table 5 Most used datasets for the FS task in structured or labeled data applications

Datasets & Reference	Working field
Breast cancer [47], Ovarian cancer [48], Lung cancer [47], Prostate cancer [48], Colon [47], CNS [47], SRBCT [47], ALL_MLL [47], Lymphoma [47], MLL [47], GLI_85 [49], DLBCL, 9_Tumors, 11_Tumors, 14_Tumors, Brain_Tumor1, and Brain_Tumor2 [50]	Medical data classification
Typhoons and Hurricanes(1851–2014) [51], Rain in Australia dataset [52], yeast [53], Ionosphere [54], BASEHOCK [55], Spambase [56], Kinase inhibition [57], Wine [58]	Classification task

Table 6 Description of most commonly used features associated with the FS task for the structured or labeled data applications

Feature	Description
Categorical features [59]	– These types of features represent the property of the data. It means it can split the data items into various groups. E.g., gender, age group, language, etc. It is of two types ordinal and nominal. 1. *Nominal :* Without offering any quantitative value, it is used to label the variables. This type of feature is neither measured nor ordered. 2. *Ordinal :* Ordinal data is a statistical category of quantitative data through which variables reside in ordered categories. It doesn't follow any distance measure.
Numerical features [60]	– This type of feature in which the information is in a measurable form. It is of four kinds like discrete, continuous, interval, ratio, etc. 1. *Discrete :* It is the kind of information that can only take specific values. These values don't have to be whole numbers. 2. *Continuous :* These types of features take any values. E.g., temperature, height, weight, length, etc. 3. *Interval :* It is a type of feature calculated along a scale in which each point is located from each other at an equal distance (interval). 4. *Ratio :*This feature is also quantitative data, having the same characteristics as interval, with an equivalent and definitive ratio being treated as a point of origin between each data and absolute zero.

ML approaches. In this section, we will observe the different graphical representations of the outcome of different FS techniques. After discussing different FS methods in various areas, the result can be view from the perspective of:

1. ***For the structured or labeled data***

   In Tang et al. [35] paper, different FS methods compared along with SVM classifier for various high dimensional datasets. For the SVM classifier, Fig. 7 shows the comparisons for different high dimensional datasets like Wine, Breast, Lymphoma, and Leukemia. Here y-axis represents the overall accuracy, and x-axis defines different FS methods like Five-way Joint MI (FJMI), Relief, Conditional Infomax FE (CIFE), Joint MI Maximization (JMIM), Interaction Weight based FS (IWFS), and mRMR. Here maximum accuracy 93.7% achieved on the Wine dataset for CIFE, IWFS FS methods. For the Breast dataset, maximum accuracy 97.1% achieved for FJMI, mRMR FS methods. For the Lymphoma dataset, maximum accuracy 86.2% achieved for the FJMI FS method. For the Leukemia dataset, maximum accuracy 97.6% achieved for FJMI, mRMR FS methods.
2. ***In the field of NLP***

   For the SVM classifier, in Fig. 8 shows the comparisons for different textual datasets like Reuters30, TDT2-30, RCV1-4Class, and 20Newsgroups for the various FS methods. Here y-axis represents the overall accuracy, and x-axis represents different FS methods like FJMI, Maximum Discrimination (MD), Chi-Square, JMIM, IWFS, and mRMR. Here maximum accuracy 76.3% achieved on the Reuters30 dataset for the FJMI FS method. For the TDT2-30 dataset, maximum accuracy 71.6% achieved for the FJMI FS method. For the RCV1-4Class dataset, maximum accuracy 79.7% achieved for the FJMI FS method. For the 20Newsgroups dataset, maximum accuracy 34.8% achieved for the Chi-Square FS method.
3. ***In the field of image processing***

   For the SVM classifier, Fig. 9 shows the comparisons of different hyperspectral image classification for the datasets like Botswana, KSC, and Indian Pines against various FS methods. Here y-axis represents the overall accuracy, and the x-axis represents different FS methods like mRMR, CMIM, JMI, Relief, DE, GWO, and ALO. Here maximum accuracy 93.39% achieved on the Botswana dataset for the ALO FS method. For the KSC dataset, maximum accuracy 92.77% achieved for the GWO FS method. For the Indian Pines dataset, maximum accuracy 85.52% achieved for the ALO FS method.
4. ***In the field of speech processing***

   Özseven et al. [175] paper shows this comparison for multiple speech datasets. In Fig. 10 for different speech datasets like EMOVO, eNTERFACE05, EMO-DB, and SAVEE result comparison is computed for different

Table 7 Various kinds of FS work in the area of structured or labeled data applications

Author's & references	FS Method	Datasets & performance measures	Purpose of work
Wahid et al. [61]	IG with Correlation Based	– The datasets leukemia, lung cancer, Prostate, Colon, and DLBCL were used. – For performance measures, misclassification rates, and stability analysis were used.	Diseases classification from omics data
Wang et al. [44]	BCO	– The datasets Breast Cancer Wisconsin (BCW), Central Nervous System (CNS), Colon, Leukemia (ALL-AML), 9-Tumors, 11-Tumors, 14-Tumors, Brain-Tumor1, Brain-Tumor2, SRBCT, Leukemia1, Leukemia2, Prostate-Tumor, Lung-Cancer, and DLBCL used. – The performance measures classification accuracy and the average classification accuracy rate used.	Finding cancer from gene expression of data
Potharaju et al. [62]	IG, Entropy with Correlation-based	– The datasets COLON, LEUKEMIA, LEUKEMIA-3C, LEUKEMIA-4C, LYMPHOMA, SRBT, MLL, and MUSK used, taken from http://csse.szu.edu.cn/staff/zhuzx/Datase-ts.html – The performance measures accuracy and RMS error rate used.	Genes classification from autism data
Sun et al. [63]	Entropy with F-Score	– The datasets Brain-Tumor2, Colon, DLBCL, Leukemia, Leukemia1, Lung, Prostate, Prostate1, SRBCT, and 9-Tumors used. – The performance measures reduction rate and accuracy used. For statistical significance, the Friedman test and the Bonferroni-Dunn test were used.	Finding cancer from gene expression of data
Arora et al. [64]	Chi-Square, Gain Ratio, Relief, LASSO	– The dataset German Credit, Lending Club's, and Kaggle's Bank Loan Status used. – The performance measures Jaccard Stability Measure (JSM), accuracy, and AUC used.	Classify credit risk analysis.
Jain et al. [65]	Correlation, PSO based	– The datasets Colon, Tumor, Leukemia 2-class (ALL-AML), Leukemia 3-class (ALL-AML-3), Leukemia 4-class (ALL-AML-4), CNS, Breast, Lung, Ovarian, Lymphoma, MLL, small round blue cell tumors (SRBCT) used. – The performance measures accuracy, mean classification accuracy, and the average number of genes selected used.	Finding of cancer.
Moon et al. [66]	CBFS	– The dataset was collected from ECMWF and RDAPS (which is composed of 22 different South Korea locations in the year 2013 to 2014). – Here Heidke Skill Score and Peirce skill score are used for performance evaluation.	Weather forecasting
S.B. et al. [67]	CBFS	– The datasets Atlantic hurricane database and Pacific hurricane database used, taken from the Hurricanes and Typhoons, 1851–2014 database. – The performance measures FS accuracy, Prediction Accuracy (PA), FPR, and Prediction Time(PT) used.	Weather forecasting

Table 7 (continued)

Author's & references	FS Method	Datasets & performance measures	Purpose of work
Wang et al. [68]	BCO	– The datasets 9-Tumors, 11-Tumors, 14-Tumors, Brain-Tumor1, Brain-Tumor2, SRBCT, Leukemia1, Leukemia2, Prostate-Tumor, Lung-Cancer 1, and DLBCL used. – The performance measures classification accuracy and the average accuracy rate used.	Finding cancer from gene expression of data

Table 8 Most commonly adopted FS methods in structured or labeled data applications with their advantages

FS method	Advantages
Relief based	– It has advantages over the features having two-way interaction and univariate effects [69]. – It is efficient for space where we have allocated low weights to features that are less likely to be important [70]. – Without any assumption of the size of data or population size, it selects the features. It evaluates the features in the context of the rest features and not parametric in the evaluation case [69].
Evolutionary method based	– These methods don't require any assumption about search space and the domain knowledge, such that whether it is linearly or non-linearly distinguishable in the search for optimal feature subset [71]. – As these mechanisms are based upon population, hence it can generate various solutions in a single run [71]. – As these methods are capable of multiobjective FS, hence it can produce the non-dominated feature subsets [71].
Correlation based	– It uses the best first search to evaluates the search space to get the optimal feature subset. It maintains the priority queue to preserve the solution space [72]. – Instead of evaluating the single feature, it goes for subset evaluation to generate the optimal feature subset [73]. – It is capable of selecting the features of highly correlated and lower redundant through the classes [72].

Table 9 Most commonly used datasets for the FS task in NLP applications

Datasets & Reference	Working field
Yelp [74], Twitter [75], Amazon product reviews [76], Cornell Movie Review [77], SemEval-2014 shared task [78]	Sentiment analysis
Reuters-21578 [79], TDT2 [79], WebKB [80], 20-Newsgroups “bydate” version [81], TecTc100 [82]	Text classification

Table 10 The description of various types of features used in the FS task for NLP applications

Feature	Description
Term Frequency (tf) [83]	– It is defined as how frequently in a document a term appears. Terms correspond to words or phrases in the context of NLP.
Inverse Document Frequency (idf) [84]	– IDF is a statistical measure that calculates the significance of a term in a text document collection. If the word appears more in the documents, its idf score will be lower. The lesser score indicates that the word is less important. e.g., a, the, etc.
$tf - idf$ [85]	– A statistical score calculates how important a word is to a document in a collection of documents. It is done by multiplying tf with idf.
Sentence Length [86]	– It defines the corresponding length of the sentence. A sentence length below the predefined threshold value is neglected automatically.
Stop word [87]	– The most used or common words in any natural language are known as stop words. For the analysis of textual data and creating the NLP models, these words do not add much value to the document's context. e.g., to, at, where, when, is, in, the, a, for, etc.
Title word [86]	– Sentences that include words that appear in the title, suggest the central theme of the document. The total frequency of keywords in the title is taken as a specific feature.
Biased word [86]	– If the sentence contains one word or more from a biased word list (i.e., a predefined list that may include domain-specific words), then the sentence is relevant.
Keywords [86]	– Keywords are the most relevant, essential words of the document.
Prefix and suffix [87]	– Prefix and suffix of the document can be used as an essential feature for the model. It is evaluated as a fixed-length sequence of characters.
n-grams Character [87]	– n-gram character used as a feature for the model. It is a sequence of n characters that could be extracted from the words of the sentence.
Part-of-Speech (PoS) tag [87]	– The method of assignment of one part of speech to a corresponding word is called PoS tagging. It is one of the essential features in the classification of text.
Head word PoS [87]	– The PoS of the head-word is used as a feature of a specific model.
Named entity information [87]	– Named entity belongs to the word is identified by predefined categories like organization, place, person, etc.
Semantic orientation score [88]	– Sentiment orientation score is calculated for each word, which tells how much word is associated with positive or negative sentiments.
Code quality principle [83]	– A high number of units (i.e., noun phrases or words) convey the most relevant information within a text. The Code quality principle proves a proportional relationship between how important the data is and the number of coding elements it has.
WordNet [89]	– Different words that are semantically identical (or synonymous with each other) are classified into WordNet synsets.

Table 10 (continued)

Feature	Description
Concept Similarity of a Sentence (CSS) [90]	– CSS is the number of synonym sets of query terms corresponding to the words in the sentence. The synonym sets are drawn from the WordNet to assign CSS value.
Sentiments (Emotions described by Text) [91]	– Sentiments attached in the text are known as the semantic features of the text. There are various types of sentiments attached in the text like fear, disgust, joy, hate, fear, positive, negative, etc.
Dependency relationship [92]	– Dependency relationships express the grammatical relationships within the words in the sentence.
Sentence Reference Index (SRI) [90]	– A sentence that precedes a sentence that includes a pronominal reference is given more weight by SRI.

Table 11 Various kinds of FS work in the area of NLP

Author's & references	FS Method	Datasets & performance measures	Purpose of work
Bharti et al. [93]	PSO	– The datasets Reuters-21,578, Classic4, and WebKB were used. – The performance measures F1-measure, precision, and recall were used.	Classification of textual documents
Baccianella et al. [94]	IG based	– The datasets TripAdvisor-15763, Amazon-83713 used, adapted from http://hlt.isti.cnr.it/reviewdata/. – The performance measure used here, macro-averaged mean absolute error (MAE^M) measure, where M superscript indicates macro averaging.	Classification of textual documents.
Yousefpour et al. [95]	Chi-Square, IG	– The datasets movie review, book review, electronic review, kitchen review, and music review used, taken from https://www.cs.jhu.edu/~mdredze/datasets/sent-iment/ and https://www.cs.cornell.edu/people/pabo/movie-review-data/. – The performance measures used here recall(neg), recall(pos), precision(neg), precision(pos), F1- score, accuracy.	Identification of textual sentiment analysis.
Sundararaman et al. [96]	IG based	– The dataset Congestive Heart Failure (CHF) was used, adapted from the MIMIC III database. – The performance measures used here recall(neg), recall(pos), precision(neg), precision(pos), F1-Score, accuracy.	Prediction of hospital readmission
Kushwaha et al. [43]	PSO	– The datasets Reuters-21578, TDT2 and TR11 used, taken from http://www.cad.zju.edu.cn/home/dengcai/-Data/TextData.html. – The performance measures accuracy, purity, rand-index, and NMI used.	Classification of textual documents

Table 11 (continued)

Author's & references	FS Method	Datasets & performance measures	Purpose of work
Manochand-ar et al. [97]	$tf-idf$ based	– The datasets Mobile, Movie(MR1), Yelp, IMDb, MR2, CR, SemEval, Pros & Cons, and Subj used. – The performance measures accuracy, precision, recall, F-Measure, True Negative Rate (TNR), False Positive Rate (FPR), False Negative Rate (FNR), False Discovery Rate (FDR), Negative Predictive Value (NPV), effectiveness measure used. Also, Strength of the Expressions(SE(i)), Selection of Weightage values, and Statistical significance of t-test for SE(i) used.	Classification of public reviews
Al-Salemi et al. [98]	MI based	– The datasets Reuters-21578, 20-Newsgroups, OHSUMED, TMC2007 used. – The performance measures for classification macro-averaged F1 (MacroF1) and micro-averaged F1 (MicroF1). Friedman test for the evaluation of the boosting algorithm.	Classification of textual documents
Yarlagadda et al. [99]	$tf-idf$ based	– The datasets 20-Newsgroups, and Reuter used. – The performance measures F1-measure, precision, and recall used.	Classification of textual documents
Rehman et al. [100]	PSO	– The datasets WAP, K1a, K1b,re0, and re1 were used, taken from Karypis Lab, University of Minnesota, and Concept drift adopted. The dataset 20Newsgroups also used. – The performance measures MacroF1 and MicroF1 used.	Classification of textual documents
Zhang et al. [101]	PSO based	– Spam dataset used, collected from UCI repository in the year 1999. Manually created dataset used for the year 2012. – The performance measures sensitivity, specificity, classification accuracy rate, and confusion matrix used.	Detection of spam
Sanghani et al. [102]	tf based	– The three datasets ENRON, ECML, PU were used. – The performance measures sensitivity, specificity, classification accuracy rate, and confusion matrix used.	Development of e-mail spam filter
Metin et al. [103]	IG, GR, Chi-Square, Relief	– Six different Turkish corpora are BilCol, Bilkent, Ege, Leipzig, Metu, and Muder corpus. – The performance measures precision, recall, and F1-values used.	Identification of Turkish Multiword expression

FS methods like SFS, OM(th_{SD}), OM (th_{MN}), OM (th_{MED}), and OM (th_{CV}) against various classifiers like SVM, MLP, and k-NN. Here maximum accuracy 85.71% achieved with SFS and OM (th_{MN}) FS method via SVM classifier on the EMO-DB dataset. For the eNTERFACE05 dataset, maximum accuracy 68.46% achieved with OM (th_{MN}) FS method via MLP classifier. For the SAVEE dataset, maximum accuracy 77.92% achieved with OM (th_{MN}) FS method via SVM classifier. For the EMOVO dataset, maximum accuracy 63.91% achieved with the OM (th_{SD}) FS method via SVM classifier.

Table 12 Most commonly adopted FS methods in NLP applications with their advantages

FS method	Advantages
PSO .	– PSO can evaluate the smaller number of subsets for the large collection of terms that have been selected by the IG [104]. – The correlation coefficient can evaluate the fitness in the search for the optimal categorization or classification accuracy [104]. – Ensemble learning can be embedded in the PSO. E.g., in the paper [87] sentiment analysis task has been done. – PSO can be employed for the random search to find out the optimal number of feature sets. CHI method is improved via PSO [105]. – CFS is employed in the PSO for better relevance and minimum redundancy in the final feature subset [106]. – Instead of using the fixed-parameter update, PSO can be applicable with some adaptive parameter updates [107].
Term weighting approach .	– Suppose the term appears hardly in a particular class and doesn't appear in rest classes. Then it is not a relevant term and has been assigned with a lower score. Otherwise, it will get a higher score [108]. – Suppose the term appears regularly in a particular class and doesn't appear in rest classes. Then it is a relevant term and has been assigned with a higher score. Otherwise, it will get a lower score [108]. – It can able to differentiate between the positive and negative documents. It also enhanced the power of discrimination of terms for text classification [109]. – It can handle the shortcoming of classes with a smaller number of instances are discriminated, and the corresponding resultant classifier performs below the acceptable result [110]. – These approaches can measure the terms that have the class discriminant capability [111].
IG .	– IG able to calculate the number of bits of the information absorbed for the prediction of category. The selected features belong to the most diverse capability [112]. – IG evaluates the most important terms by discarding the outliers from the overall corpus [113]. – It calculates the features that can be the most informative with the document levels [114]. – IG can handle the situation where the features too redundant because the documents in which they are separable are extremely overlapping in nature [115]. – It gives the best results on imbalanced datasets and highly skewed datasets [113].
t-test .	– t-test can be able to handle the heterogeneity for the distribution of tf among the corpus and a particular category [116]. – the t-test can be able to filter out the minimum frequency term corresponding to their minimum weight [117]. – It can figure out the relation among the tf and the category topics [117]. – In topic modeling, a t-test can be applied for the multinomial distribution of words [118]. – t-test can test that the average tf of a specific term among the two classes are different statistically [119].

Table 13 Most commonly used datasets for the FS task in image processing applications

Datasets & Reference	Working field
Facial Recognition Technology (FERET) [120], extended Yale Face Database B [121], Olivetti Research Laboratory(ORL) [122], CASIA database [123], LFW face database [124], PIE [125], JAFFE [126]	Biometrics
Kennedy Space Center (KSC) [127], Reflective Optics System Imaging Spectrometer (ROSIS-03) [127], Washington DC Mall [128], Indian Pines [128], Botswana [129]	Image classification
National Cancer Institute database [130], Repository of Kent Ridge Biomedical Dataset [131], LL_SUB and GLI_85 [132], LUNA16 [133], Kaggle MRI dataset [134], STARE [135]	Medical image classification

Table 14 The description of various types of features used in the FS task for image processing applications

Feature	Description
Geometrical features	– These features describe the geometric properties of images, e.g., surfaces, curves, points, lines, etc. e.g., *blobs, corners, ridges, edges, image texture based feature, silent point based features, etc.*
frequency and space domain descriptor	– Some of the features are also derived from the frequency and space domains. Almost these features are constructed from the FT and the GT. *e.g., Discrete Cosine Transform (DCT), Gabor Transform (GT), Wavelet Transform (WT), Scale Invariant Feature Transform(SIFT), oriented fast and rotated brief (ORB), Speeded-Up Robust Features (SURF), Binary Robust Invariant Scalable Keypoints (BRISK), Binary Robust Independent Elementary Feature (BRIEF), Maximally Stable Extremal Regions (MSER), etc.* 1. *DCT:* DCT tells the definite sequence of data points in terms of summation of cos function at various frequencies. DCT based features are used for face recognition [136]. 2. *GT:* A particular case of the short-time FT is the GT, named after Dennis Gabor. As it varies over time, it is used to determine the sinusoidal frequency and phase content of local parts of a signal. This feature is used to classify handwritten characters [137]. 3. *WT based features:* Fixed window size is added to the GT, which can be one of its limitations since certain textures can be characterized according to different scales. The WT is based on multiscale analysis of the images and uses observation windows of different sizes in order to resolve this limit. These features are used for the classification of texture images [138]. 4. *SIFT:* It is used to achieve lighting robustness. By encoding the image, variations, and small positional shifts in a localized collection of gradient orientation knowledge with histograms [139]. 5. *SURF:* This is the alternate version of SIFT. Its combination of the Hessian-Laplace region detector and its feature descriptor of gradient-based [140]. 6. *BRIEF:* It performs a binary test among the pixels in smoother image regions. It is very compassionate in the plane rotation [141]. 7. *ORB:* It is the best alternate for SURF AND BRIEF. It has an advantage over noise and rotation invariant [142]. 8. *MSER:* For the detection of the blob, this feature is useful [142].
Color & shape descriptor based	– For image representation, color is an essential feature, it is invariant in terms of transition, scaling, and rotation of the image [143]. For the similarity matching, the shape is the crucial feature of the image [144]. *e.g. Histogram based, Color coherent vector based, Color moments based, 3-D shape based, Region based, Contour-Based shape, Contour Based, etc.*
Texture descriptor based	– These features are used to divide images into regions of interest and to categorize those areas. Texture offers details about the spatial arrangement of an image's color or intensity. *e.g. Haralick features, Histograms of Oriented Gradients (Hog), Gray Level Cooccurrence Matrix (GLCM), Autoregressive Model Features (ARM), Run Length Matrix (RLM), Local Binary Patterns (LBP), etc.* 1. *Haralick feature:* It collects information about the trends that emerge in texture patterns [142]. 2. *Hog:* It helps to differentiate how intensity gradients are distributed [145]. 3. *GLCM:* It is used for the extraction of the textual feature with their various orientations of an image. It depends upon the spatial relationships among the pixels [142]. 4. *ARM:* These features treat an image as a pixel sequence and describe its likelihood as the product of all pixels' conditional probabilities [146]. 5. *RLM:* This matrix of features is created via the run length of a grayscale image. It is defined as the collinear, consecutive pixels of an image having the same gray level [146]. 6. *LBP:* It is used to distinguish the textual area of the image by comparing it with the level of luminance of each pixel to its neighboring pixels [147].

Table 15 Various kinds of FS work in the area of image processing

Author's & references	FS Method	Datasets & performance measures	Purpose of work
Vignolo et al. [148]	MI with GA	– The dataset Essex Face Database used, adopted from the Vision group. – The performance measure accuracy, ROC curves, Relative Error Reduction (RER), and convergence were used.	Facial recognition system
Togaçar et al. [37]	mRMR	– The dataset LeNet, AlexNet, and VGG-16 used. – The performance measures accuracy, sensitivity and specificity, precision, and F-score used.	Detect Lung cancer from CT images
Sharma et al. [149]	MI	– The dataset was created via 90 ultrasound liver images collected from Delta Diagnostic Centre Patiala, India. – The performance measures sensitivity rate, specificity rate, accuracy rate, Miss Rate, and AUC curve used.	Detect fatty liver diseases from ultrasonography images
Wang et al. [150]	PSO	– The dataset used is of 40MHz catheter probe images. – The performance measures used here are Jaccard, Hausdorff Distance (HD), and Percentage of Area Difference (PAD).	Detection of Media-adventitia border
Shao et al. [38]	Correlation based	– The dataset Human Protein Atlas (HPA) used. – The performance measures subset accuracy, classification accuracy, and Recall used.	Classification of Multi-Label Subcellular Bio-Images
Yurtkan et al. [151]	Entropy based	– The dataset BU-3DFE is used here. – The performance measures used recognition rate and classification rates.	Classification of facial expression
Krisshnaa et al. [152]	PSO	– The datasets ORL, University of Manchester Institute of Science and Technology (UMIST), Extended YaleB, Carnegie Mellon University Pose, Illumination, and Expression (CMUPIE), FERET, Foundation for Education of Ignatius (FEI), and Pointing Head Pose Image Database (HP) used. – The performance measures Training time, Testing time, Average recognition rate in (%), and Maximum recognition rate in (%) used.	Facial recognition system
Mistry et al. [153]	GA with PSO	– The datasets CK+ and MMI were used. – The performance measures recognition rates and classification rates used.	Classification of facial emotions
Peralta et al. [154]	IG based	– The datasets SFinGe, NIST-SD4, and NIST-SD14 were used. – The performance measures classification accuracy, rejection rates, ideal penetration rate, optimal penetration rate, TPR, FPR, and average identification time rate used.	Classification of fingerprints

Table 16 Most commonly adopted FS methods in image processing applications with their advantages

FS method	Advantages
MI based	– It gives the best result to assign the rank of image features having various wavelengths [155]. – It can provide various advantages like non-linearity, free of distribution, and lower computational complexity for multiclass cases in image domains [156]. – It can evaluate the optimum measurement for the complementation between the image features [156]. – It can easily handle the situation when the images are harder to discriminate from each other because the bands are very similar [156]. – It can handle the situation when the entropy is not bounded in [0,1] [156].
PSO	– DCT gives an outstanding outcome for the compaction of energy. PSO can able to handle the DCT results due to the distance measure it uses [157]. – PSO works efficiently in the search spaces of images that are irregular, noisy, and frequently change over time [158]. – For image processing, PSO can be flexible to use the scattered index as its fitness function via histogram [157]. – PSO can efficiently work on the edge information, standard deviation, mean values of pixels, and entropy of the image [159]. – PSO can able to embedded in mesh optimization for image processing problems [160].
Dependency measure based	– [–] It can handle the various orientation of features. As in image problems, it gives the best results. – It can handle the indiscernibility of features more efficiently. – It gives the best results for medical image datasets when there are cases when the boundary values are unclear.
Probablity based	– It can handle features like various image regions, i.e., local (may have different characteristics), uniform, non-uniform, etc [161]. – It can perform well to differentiate in texture and non-textured features [162]. – It can able to categorize the features when the luminance of pixels are distributed evenly [161].

7 Challenges and future scope

After the detailed analysis of the FS tasks in the various field of ML in this section, we have highlighted the various issues and the future scope of the FS task. We have divided the issues and future scope for the structured or labeled data and the unstructured data.

7.1 In structured or labeled data

There are several challenges hits when dealing with structured or labeled data for the FS method, they are:

- Feedback in real-time is of vital importance; this means that online FS methods are needed, which is still a challenge for researchers.
- Another prevalent challenge in microarray data is a class imbalance. This condition occurs when, due to more instances in the data than the other minority classes, a dataset is dominated by a major class.
- As a result of splitting original datasets into training and test sets, the dataset change takes place. This occurs when there is a situation in which the class distribution of inputs and outputs differs between training phases and test phases in the holdout (test set) results. As a consequence, in real-world settings, the traditional notion that the training and test data contain similar distributions is typically violated, which may hamper the gene selection and classification process.

Some of the future scope for the FS task when dealing with the structured or labeled data, which are:

- This analysis of the methods of FS shows that a specific algorithm for FS plays a vital role in the accurate classification of diseases. Therefore, it is important to apply more applications and improvements to such methods in disease datasets.

Table 17 Most commonly used datasets for the FS task in speech processing applications

Datasets & Reference	Working field
IEMOCAP [163], EMO-DB [164], SAVEE [165], CASIA [166], eNTERFACE'05 [167], EMOVO [168], MASC [169]	Speech Emotion Recognition

Table 18 The description of various types of features used in the FS task for speech processing applications

Feature	Description
Prosodic features	– These features occurred in a connected speech when we bring speech or signal together [170]. Human beings perceive these features in terms of intonation and rhythm. These features are also known as paralinguistic features because they deal with the phrases, words, syllables, words, of the speech signal.*e.g. pitch etc.* 1. *Pitch :* It is termed as a degree of loudness and hardness of voice [171].
Temporal features (Time domain based)	– These features represent the long-term dynamics of a speech signal over time [172].*e.g., tempo, minimum energy, zero-crossing rate, short-time energy, the energy of the signal, maximum amplitude, etc..* 1. *Short time energy :* It is a useful feature to differentiate the voice and non-voice segments. It is also helpful to detect the ending point of utterance [171]. 2. *Tempo :* Each speech signal has its speed, which has been measured by the feature tempo. It is usually measured beats per minute [171]. 3. *Zero crossing rate :* It is used for the depiction to concentrate the energy in the spectrum. It is also helpful to differentiate the speech from the noise [171].
Spectral features (Frequency-based)	– These types of features are obtained by converting the time domain based speech signal into the frequency domain based speech signal via Fourier transform [170].*e.g. spectral roll-off, spectral centroid, fundamental frequency, spectral flux, frequency components, spectral density* [173], *Mel Frequency Cepstral Coefficients (MFCC), Linear Prediction Cepstral Coefficients(LPCC), Log-Frequency Power Coefficients (LFPC), Gammatone Frequency Cepstral Coefficients (GFCC).* 1. *Spectral flux :* This feature is used to separate the music from the speech signal. It measures how rapidly the power spectrum of a speech signal changes. It is also used for the determination of timber from the audio signal [171]. 2. *MFCC :* This feature demonstrates the short-term power spectrum of an audio signal [171]. 3. *LPCC :* This feature describes the vocal tract properties of the speaker [171]. 4. *LPCC :* This feature extracts the logarithmic filter property of the human voice [171]. 5. *GFCC :* This feature is also obtained by the MFCC method. In the power spectrum, instead of the mel filter bank, the gammatone filter bank has been used [171].
Voice quality features	– With the help of the vocal tract's physical characteristics, voice quality features have been calculated [170].*e.g., Jitter, Normalized amplitude quotient (NAQ), Shimmer, Quasi open quotient (QOQ), harmonics to noise ratio (HNR)* 1. *Jitter :* It is defined as the variation of frequency from one cycle to another. It measures the instability in frequency [174]. 2. *Shimmer :* It is defined as the variation of amplitude in the sound wave. It measures the instability in amplitude [174]. 3. *HNR:* This feature specifies the amount of additional noise in the speech signal. [174]. 4. *NAQ:* Using two amplitude-domain measurements from waveforms calculated by inverse filtering, the NAQ is provided as a tool to parameterize the glottal closing step [174]. 5. *QOQ:* The QOQ is a commonly used open quotient correlation involving the derivation of the quasi-open phase based on glottal amplitude measurements [174].
Teager energy based features	– Some of the features have been generated by the Teager energy operator. It is used to identify the stress in the speech signal [170].

Table 19 Various kinds of FS work in the area of speech processing

Author's & references	FS Method	Datasets & performance measures	Purpose of work
Özseven et al. [175]	Proposed mean-threshold based FS	– The datasets EMO-DB, eNTERFACE05, EMOVO, and SAVEE were used. – The performance measure classification accuracy(%) and total workload used.	Classification of emotions from the speech signal
Casale et al. [176]	GA	– The dataset Speech Under Simulated and Actual Stress (SUSAS) was used. – The performance measures the detection rate and classification rate used.	Classification of emotions from the speech signal
Liu et al. [40]	Correlation with F-Score	– The dataset CASIA from the Institute of automation of the Chinese academy of sciences was used. – The performance measure recognition rate used.	Classification of emotions from the speech signal
Mirzaei et al. [21]	MI based	– The dataset used data provided by AP-HP Broca Hospital in Paris. – The performance measures classification accuracy and confusion matrix used.	Detect Alzheimer's disease from the human voice
Mencattini et al. [177]	Correlation based	– The dataset CASIA from the Institute of automation of the Chinese academy of sciences was used. – The performance measure coefficient of determination(R^2), Residual Sum of Squares (RSS), Total Sum of Square (TSS), ROC, AUC, and Percentage of selected features used.	Classification of emotions from the speech signal
Rong et al. [178]	Probability with C4.5 DT and RF ensemble	– The datasets created via the acted speech corpora and natural speech corpora both are Chinese (Mandarin). – The performance measure classification accuracy and confusion matrix used.	Classification of emotions from the speech signal
Saari et al. [20]	Correlation based	– The dataset adopted from https://www.jyu.fi/music/coe/materials/emotion/-soundtracks. – The performance measures classification accuracy, precision, recall, F-measure, and confusion matrix used.	Classify expression of mood in the musical audio signal

- Future research based on labeled data is likely to benefit from this fruitful field of FS is novel collective or ensemble feature selection methods and more robust (e.g., statistical) approaches to evaluating a selection cutoff threshold.
- There is minimal research on the use of a graph-based unsupervised FS approach for the labeled data. From a future perspective, a robust graph-based unsupervised FS approach is needed for the labeled data.

7.2 In unstructured data

Here we will sketch the issues and the FS method's future scope in NLP, image processing, and speech processing.

7.2.1 In NLP

There are some challenges for the FS method in the context of NLP, which are listed below,

- However, the Chinese language has its unique characteristics and features (like some other Asian languages). There is no common meaning of a word, for instance, and there are no written Chinese spaces between characters. This is the biggest challenge for the FS task in the text classification problem.
- A text document is a set of words arranged according to the grammatical rules of their respective language. Although many term weighting schemes exist for TC,

Table 20 Most commonly adopted FS methods in speech processing applications with their advantages

FS method	Advantages
Correlation-based	– It is very efficient to discard the emotional features of the speech [179]. – In speech features, there exists a mutual reinforce or mutual restraint relationship. This approach can be able to control these types of situations [179]. – Spearman rank correlation coefficient measure can able to find out the relation between the speech features under the monotonic condition [179].
F-score based	– It can discard the effect of between-class variation, which can distort the association of feature subsets in the corresponding distributed class [179]. – It can eliminate the problems in which PCA can't extract the discriminated information from the emotional features of the speech signal [179]. – It performs well to differentiate the emotion of the speech.

it remains a significant challenge to find an appropriate term weighting scheme.

- Named Entity Recognition (NER) is the critical step for the FS task in text mining. Some languages or domains are low in resources, thereby making the NER role very difficult. The change in the tag set increases the system's complexity. For more organizations to be known, more laws or characteristics should be defined.
- It is still very challenging to choose the optimum feature extraction technique for underlying multi-modal data. The overall model quality and reliability are essentially influenced by how effectively its feature vectors have portrayed a modality.

The future scope for the FS method in the context of NLP, which are listed below,

- Many new approaches to dealing with linguistic features increase the quality of the NER method. Still, they need more linguistic expertise and complicated linguistic techniques at the expense of more computational time and memory space. Therefore, statistical NER Systems need to be built that can identify relevant, well-quality named entities. So, that the FS task performance can be enhanced.
- The efficient cross-modal methods that can be implemented in a distributed environment are required to enhance the FS task's performance. More research is needed to develop successful cross-modal algorithms that can be applied to enormous multi-modal datasets.
- Further research works will focus on establishing the FE method, which will be based on vocabulary, syntactic, semantic, and ontology to enhance the FS task's performance.
- The core part FS task is the objective function. The researchers are still in progress trying to find the optimal objective function in the text mining field.

7.2.2 In image processing

There are some challenges for the FS method in the context of image processing, which is listed below,

- The challenges of image processing tasks mostly revolve around high-resolution images, including the effects of exponential dimension growth on the choice of FS process, contextual elements that affect FS output, and the existence of noise in different types.

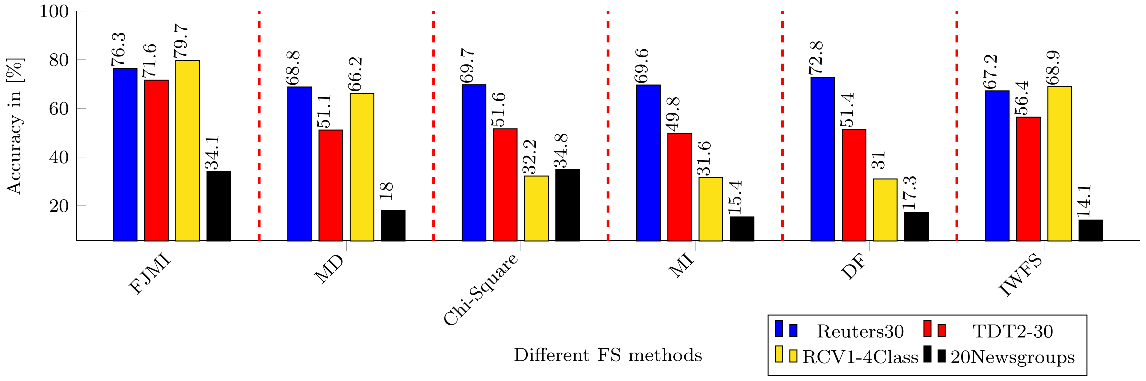

Fig. 7 Comparison of accuracy for the classification of structured or labeled data of various labeled datasets against the different FS approaches

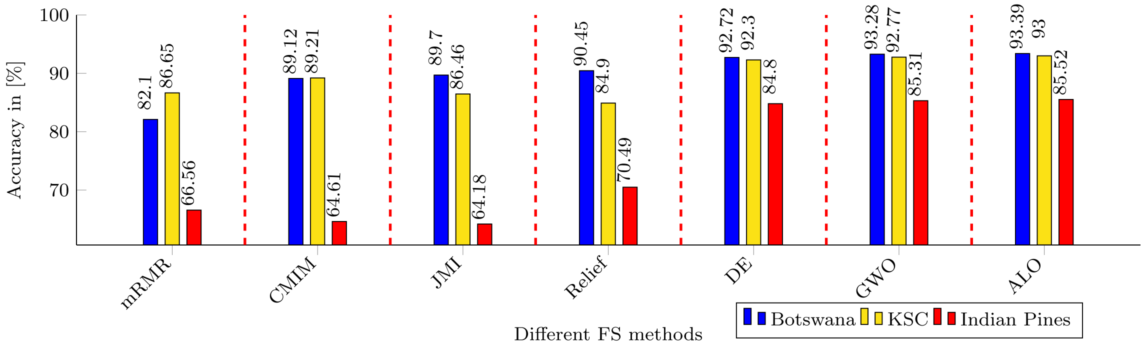

Fig. 8 Comparison of accuracy for the text classification of various textual datasets against the different FS approaches

- [–] Another factor that limits the FS task is object orientation, affecting the quality of video and image data. If a subject's head moves in 3D space, then the task of identifying facial expression becomes difficult. Item orientations (context of text), such as font size, patterns, alignment, and color, make it challenging to detect text in video images.
- The major challenge for unsupervised FS has remained to find the required evaluation criterion or pseudo marks. As class labels are not available in unsupervised FS, it hasn't proven easy to extract the related features simultaneously. Pseudo labels have been developed in recent unsupervised FS progress to be used with sparse projection to direct the FS system. However, the pseudo labels developed are noisy, ineffective, and suffer from complex multi-limit composition and computational inefficiency. On the other hand, there is no clear requirement for the proportion of labeled data to unlabelled data to be used for semi-supervised learning instances.

The future scope for the FS method in the context of image processing, which is listed below,

- [–] There is minimal number of research done for image enhancement for the FS process. To improve FS task efficiency in image processing, the objective

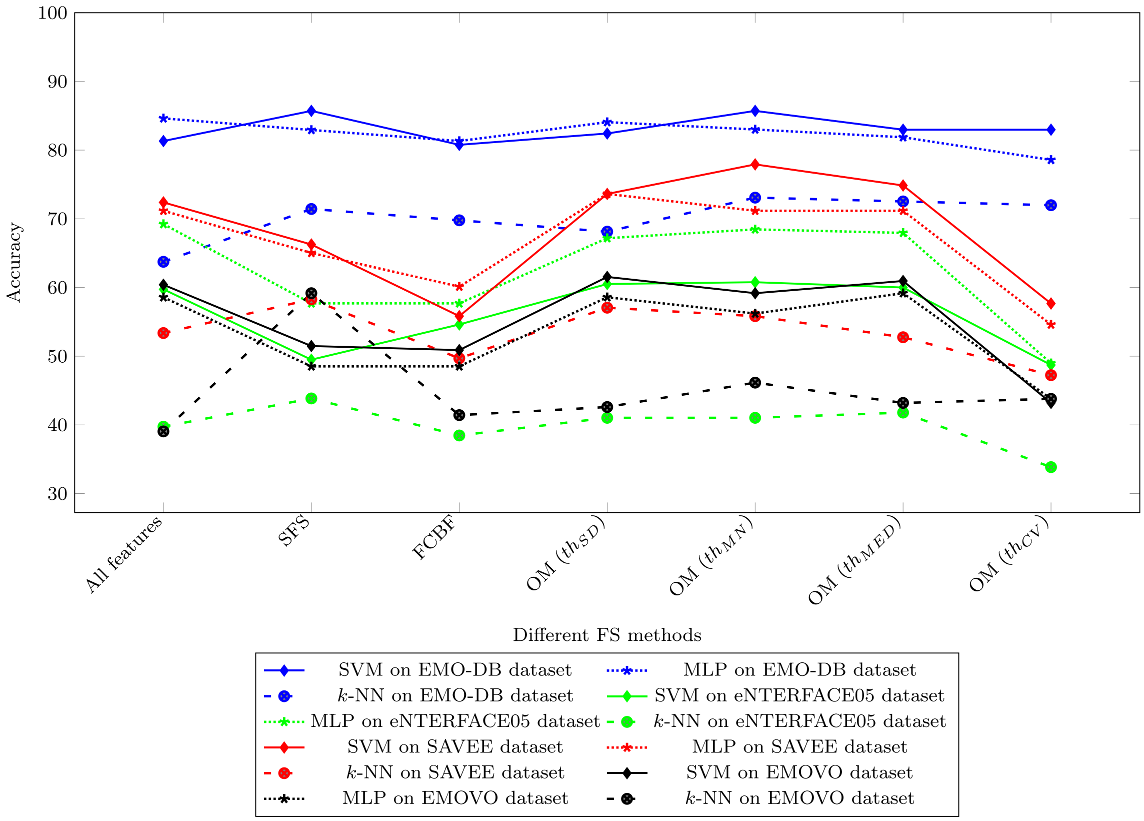

Fig. 9 Comparison of accuracy for the classification of various hyperspectral image datasets against the various FS approaches

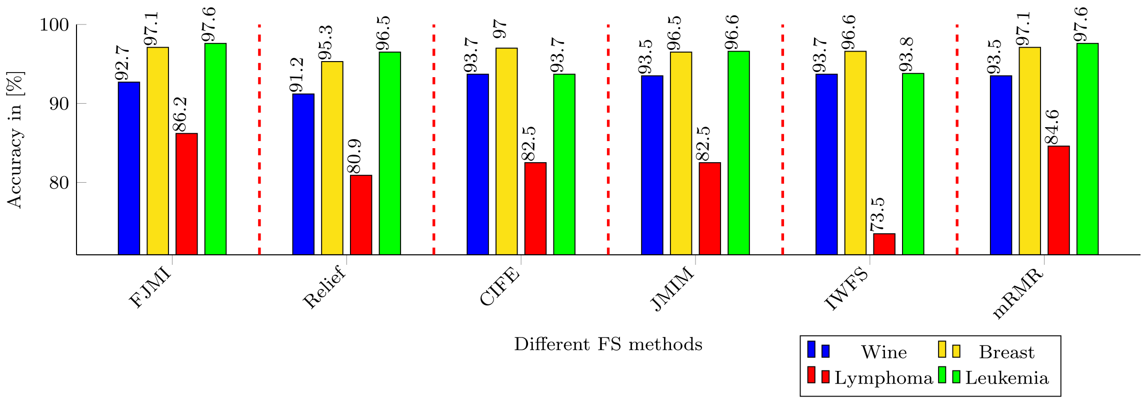

Fig. 10 Comparison of accuracy for the classification of speech recognition of various speech datasets against the different FS approaches

here is to apply super-resolution algorithms or 3-D image generation algorithms to low-resolution images. To reduce the difference between low-resolution and high-resolution images, new algorithms are needed for resolution-robust FS methods.

- About the practical applications, a formal analysis of the human function in image processing is promising. It should not be restricted to the particular aspects referred to in the studies, comprehensibility of findings, including contextual data elements, and judgment of importance to provide preliminary information to supplement FS.
- It isn't straightforward to build generic heuristic FS, and alternative methods can consist of multi-level and hybrid FS to provide localized processing stages.
- Rough set based FS is widely used in many image processing tasks for its efficient performance. However, a future scope would be applied for challenging areas like geological images.
- Evolutionary-based FS is getting very popular in image processing tasks. In future research, there is a scope for emotion detection to enhance the performance; the well-established texture descriptors can be appended to the existing feature vector.

7.2.3 In speech processing

There are some challenges for the FS method in the context of speech processing, which is listed below,

- The data set generation is one of the most significant issues that are part of the learning process. Many of the dataset's acts or elicitations are recorded in unique silent spaces. Furthermore, Real-life data is noisy and has much more distinct features than real-life data. Although there are also natural data sets, there are very fewer in practice. Legal and ethical considerations are involved in documenting and utilizing natural emotions.
- Cultural and language influences on speech are also present. There are several open research working on cross-language. The research conclusions show that the features used and the current system are not enough for it. For example, the intonation of feelings about speech between different languages can show differences.
- In a controlled environment, the absence of a broad natural database is a significant challenge for transferring results on speech emotion recognition (SER).To build an emotional database is time-consuming and costly, which limits the size of the existing corpora.
- The imbalance of data is an inherent problem of natural databases. The class imbalance occurs with respect to class, length of utterance, etc.

The future scope for the FS method in the context of speech processing, which is listed below,

- In future research, it is essential to mention a more straightforward interpretation of frame selection behaviors for phonemes and the application of frame selection to a more complicated situation.
- The use of more naturalistic corpora in speech emotion recognition in future research. Naturalistic Corpora have some common types of disturbing effects like noise, low recording quality (phone calls), (city noise), etc. For performance comparison, speaker-dependent and speaker-independent emotion estimation must be used.
- As far as future research is concerned, it is particularly important whether it would be possible to design automated sampling procedures that could ensure that comparable results could be obtained under some circumstances. In the correlation-based FS analysis, given the analysis of symmetrical uncertainty in datasets, this topic becomes more important.

8 Conclusion

FS proves by the help of its results that it is not always necessary that all the features play a significant role in ML. As this is a vast area, so we elaborate here only on the core and essential concepts. Day by day, the researchers come out with numerous ideas about FS. This paper has discussed what FS is, why it is necessary, how we implement it in ML, and where FS has been applied. We have discussed three FS models, where the filter model is much faster than all the other models, but it suffers the accuracy issue. The wrapper model gives the highest accuracy, but it bears the computational time afterword embedded model is adopted to overcome the problems from the above two models. We have discussed the most common datasets used in the FS task, the most common FS methods researchers usually adopt, and the common types of features in the feature sets used in various ML areas. Afterward, we covered the different FS methods in multiple areas, what are the datasets used, what kind of performance measure the author used, and the types of ML algorithms used. From the experimental results, there is a variation in the accuracy of the models on various datasets. Finally, we have addressed the issues and future scope for the FS task from various perspectives.

References

1. Sahu B, Dehuri S, Jagadev A (2018) A Study on the Relevance of Feature Selection Methods in Microarray Data. Open Bioinform J Bentham Open 11:117–139
2. Li J, Cheng K, Wang S, Morstatter F, Trevino RP, Tang J, Liu H (2017) Feature Selection: A Data Perspective. ACM Comput Surv 50
3. Liu H, Motoda H (1998) Feature Selection for Knowledge Discovery and Data Mining. Kluwer Academic Publishers, New York
4. Agarwal S, Dhyani A, Ranjan P (2019) Newton's second law based PSO for feature selection: Newtonian PSO. J Intell Fuzzy Syst 37:4923–4935
5. Narendra PM, Fukunaga K (1977) A branch and bound algorithm for feature subset selection. IEEE Trans Comput C-26:917–922
6. Khammassi C, Krichen S (2017) A GA-LR wrapper approach for feature selection in network intrusion detection. Comput Secur 70:255–277
7. Brassard G, Bratley P (1996) Feature Selection for Knowledge Discovery and Data Mining. Prentice Hall, New Jersey
8. Ververidis D, Kotropoulos C (2008) Fast and accurate sequential floating forward feature selection with the Bayes classifier applied to speech emotion recognition. Signal Process 88:2956–2970
9. Sharan RV, Moir TJ (2018) Pseudo-color cochleagram image feature and sequential feature selection for robust acoustic event recognition. Appl Acoust 140:198–204
10. Schumer M, Steiglitz K. (1968) Adaptive step size random search. IEEE Trans Autom Control 13:270–276
11. Coetzee FM (2005) Correcting the Kullback-Leibler distance for feature selection. Pattern Recogn Lett 26:1675–1683
12. Zhang J, Zhang J (2018) An Analysis of CNN Feature Extractor Based on KL Divergence. Int J Image Graph World Sci Publish Company 18
13. Lei S (2012) A Feature Selection Method Based on Information Gain and Genetic Algorithm. 2012 International Conference on Computer Science and Electronics Engineering, IEEE
14. Lee C, Lee GG (2006) Information gain and divergence based feature selection for machine learning based text categorization. Inf Process Manag 42:155–165
15. Yamada M, Jitkrittum W, Sigal L, Xing EP, Sugiyama M (2014) High-Dimensional Feature Selection by Feature-Wise Kernelized Lasso. Neural Comput 26:185–207
16. He X, Li L, Liu Y, Yu X, Meng J (2017) A Two-Stage Biomedical Event Trigger Detection Method Integrating Feature Selection and Word Embeddings. IEEE/ACM Trans Comput Biol Bioinform 15:1325–1332
17. Liu X, Ma L, Song L, Zhao Y, Zhao X, Zhou C (2014) Recognizing Common CT Imaging Signs of Lung Diseases through a New Feature Selection Method based on Fisher Criterion and Genetic Optimization. IEEE J Biomed Health Inform 19:635–647
18. Vijayanand R, Devaraj D, Kannapiran B (2018) Intrusion detection system for wireless mesh network using multiple support vector machine classifiers with genetic-algorithm-based feature selection. Comput Secur 77:304–314
19. Li B, Meng MQ-H (2012) Tumor Recognition in Wireless Capsule Endoscopy Images Using Textural Features and SVM-Based Feature Selection. IEEE Trans Inf Technol Biomed 16:323–329
20. Saari P, Eerola T, Lartillot O (2010) Generalizability and Simplicity as Criteria in Feature Selection: Application to Mood Classification in Music. IEEE Trans Audio Speech Lang Process 19:1802–1812
21. Mirzaei S, ElYacoubi M., Garcia-Salicetti S., Boudy J., Kahindo C., Cristancho-Lacroix V., Kerhervé H., Rigaud A-S (2018) Two-Stage Feature Selection of Voice Parameters for Early Alzheimer's Disease Prediction. IRBM 39:430–435
22. Novovicova J., Pudil P., Kittler J. (1996) Divergence based feature selection for multimodal class densities. IEEE Trans Pattern Anal Mach Intell 18:218–223
23. Zhang Y, Li S, Wang T, Zhang Z (2013) Divergence based feature selection for separate classes. Neurocomputing 101:32–42
24. Wang T, Li W (2017) Kernel learning and optimization with Hilbert-Schmidt independence criterion. Int J Mach Learn Cybern 9:1707–1717
25. Song L, Smola A, Gretton A, Bedo J, Borgwardt K (2012) Feature Selection via Dependence Maximization. J Mach Learn Res 3:1393–1434
26. Song L, Smola A, Gretton A, Borgwardt KM, Bedo J (2007) Supervised feature selection via dependence estimation. ICML '07: Proceedings of the 24th international conference on Machine learning, pp 823–830
27. Liu H, Setiono R (1996) A probabilistic approach to feature selection - a filter solution ICML'96: Proceedings of the Thirteenth International Conference on International Conference on Machine Learning, pp 319–327
28. Dash M, Liu H (2003) Consistency based search in feature selection. Artif Intell 151:155–176
29. Ötürk O, Aksaç A, Elsheikh A, Özyer T, Alhajj R (2013) A Consistency Based Feature Selection Method Allied with Linear SVMs for HIV-1 Protease Cleavage Site Prediction PLOS ONE 8

30. Neumann J, Schnörr C, Steidl G (2005) Combined SVM-Based Feature Selection and Classification. Mach Learn 61:129–150
31. Bahassine S, Madani A, Al-Sarem M, Kissi M (2020) Feature selection using an improved Chi-square for Arabic text classification. J King Saud Univ Comput Inf Sci 32:225–231
32. Huffman C, Sobral H, Hinojosa ET (2019) Laser-induced breakdown spectroscopy spectral feature selection to enhance classification capabilities: A t-test filter approach. Spectroch Acta Part B: Atom Spectrosc 162
33. Ravisankar P., Ravi V., Raghava Rao G., Bose I (2011) Detection of financial statement fraud and feature selection using data mining techniques. Decis Support Syst 50:491–500
34. Elssied NOF, Ibrahim O, Osman AH (2014) A Novel Feature Selection Based on One-Way ANOVA F-Test for E-Mail Spam Classification. Res J Appl Sci Eng Technol 7:625–638
35. Tang X, Dai Y, Xiang Y (2019) Feature selection based on feature interactions with application to text categorization. Expert Syst Appl 120:207–216
36. Azam N, Yao JT (2012) Comparison of term frequency and document frequency based feature selection metrics in text categorization. Expert Syst Appl 39:4760–4768
37. Mesut Togaçar M, Ergen B, Cömert Z (2020) Detection of lung cancer on chest CT images using minimum redundancy maximum relevance feature selection method with convolutional neural networks. Biocybern Biomed Eng 40:23–39
38. Shao W, Liu M, Xu Y-Y, Shen H-B, Zhang D (2017) An Organelle Correlation-Guided Feature Selection Approach for Classifying Multi-Label Subcellular Bio-Images. IEEE/ACM Trans Comput Biol Bioinform 15:828–838
39. Song QJ, Jiang H, Liu J (2017) Feature selection based on FDA and F-score for multi-class classification. Expert Syst Appl 81:22–27
40. Liu Z-T, Wu M, Cao W-H, Mao J-W, Xu J-P, Tan G-Z (2018) Speech emotion recognition based on feature selection and extreme learning machine decision tree. Neurocomputing 273:271–280
41. Lu W, Li Z, Chu J (2017) A novel computer-aided diagnosis system for breast MRI based on feature selection and ensemble learning. Comput Biol Med 83:157–165
42. Zini L, Noceti N, Fusco G, Odone F (2015) Structured multi-class feature selection with an application to face recognition. Pattern Recogn Lett 55:35–41
43. Kushwaha N, Pant M (2018) Link based BPSO for feature selection in big data text clustering. Fut Gener Comput Syst 82:190–199
44. Wang H, Jing K, Niu B (2017) A Discrete Bacterial Algorithm for Feature Selection in Classification of Microarray Gene Expression Cancer Data. Knowl Based Syst 126:8–19
45. Wang H, Jing X, Niu B (2016) Bacterial-inspired feature selection algorithm and its application in fault diagnosis of complex structures. In: 2016 IEEE Congress on Evolutionary Computation (CEC), pp 3809–3816
46. Kanan HR, Faez K (2008) An improved feature selection method based on ant colony optimization (ACO) evaluated on face recognition system. Appl Math Comput 205:716–725
47. http://csse.szu.edu.cn/staff/zhuzx/Datasets.html, Accessed: 2020-12-21
48. http://sdmc.i2r.a-star.edu.sg/GEDatasets/, Accessed: 2020-12-21
49. Wang M, Barbu A (2019) Are screening methods useful in feature selection? an empirical study. PLOS ONE 14(9):1–15
50. Dhal P, Azad C (2021) A multi-objective feature selection method using newton's law based pso with gwo. Appl Soft Comput 107:107394
51. https://www.kaggle.com/noaa/hurricane-database, Accessed: 2020-12-21
52. http://www.bom.gov.au/climate/data, Accessed: 2020-12-21
53. Wang Z, Wang T, Wan B, Han M (2020) Partial classifier chains with feature selection by exploiting label correlation in multi-label classification. Entropy 22(10)
54. Al-Tashi Q, Abdulkadir SJ, Rais HM, Mirjalili S, Alhussian H, Ragab MG, Alqushaibi A (2020) Binary multi-objective grey wolf optimizer for feature selection in classification. IEEE Access 8:106247–106263
55. Pilnenskiy N, Smetannikov I (2020) Feature selection algorithms as one of the python data analytical tools. Fut Internet 12(3)
56. Sakkis G, Androutsopoulos I, Paliouras G, Karkaletsis V, Spyropoulos CD, Stamatopoulos P (2001) Stacking classifiers for anti-spam filtering of e-mail. CoRR, arXiv:0106040
57. Sha Z-C, Liu Z-M, Ma C, Chen J (2021) Feature selection for multi-label classification by maximizing full-dimensional conditional mutual information. Appl Intell 51:326–340
58. Aich S, Al-Absi AA, Lee Hui K, Sain M (2019) Prediction of quality for different type of wine based on different feature sets using supervised machine learning techniques. In: 2019 21st International Conference on Advanced Communication Technology (ICACT), pp 1122–1127
59. Neves C (2014) Categorical data analysis, third edition. J Appl Stat 41(4):915–916. https://doi.org/10.1080/02664763.2013.854979
60. Maxwell JA (2010) Using numbers in qualitative research. Qual Inq 16(6):475–482
61. Wahid A, Khan DM, Iqbal N, Khan SA, Ali A, Khan M, Khan Z (2020) Feature Selection and Classification for Gene Expression Data Using Novel Correlation Based Overlapping Score Method via Chou's 5-Steps Rule. Chemometr Intell Labor Syst 199
62. Potharaju SP, Sreedevi M (2019) Distributed feature selection (DFS) strategy for microarray gene expression data to improve the classification performance. Clin Epidemiol Glob Health 7:171–176
63. Sun L, Zhang X, Qian Y, Xu J, Zhang S (2019) Feature selection using neighborhood entropy-based uncertainty measures for gene expression data classification. Inf Sci 502:18–41
64. Arora N, Kaur PD (2020) A Bolasso based consistent feature selection enabled random forest classification algorithm: An application to credit risk assessment. Appl Soft Comput 86
65. Jain I, Jain VK, Jain R (2018) Correlation Feature Selection based improved-Binary Particle Swarm Optimization for Gene Selection and Cancer Classification. Appl Soft Comput 62:203–215
66. Moon S-H, Kim Y-H (2020) An improved forecast of precipitation type using correlation based feature selection and multinomial logistic regression. Atmosph Res 40
67. Pooja S.B, Siva Balan R. V., Anisha M, Muthukumaran M. S, Jothikumar R (2020) Techniques Tanimoto correlated feature selection system and hybridization of clustering and boosting ensemble classification of remote sensed big data for weather forecasting. Comput Commun 151:266–274
68. Wang H, Tan L, Niu B (2019) Feature selection for classification of microarray gene expression cancers using Bacterial Colony Optimization with multi-dimensional population. Swarm Evol Comput 48:172–181
69. Urbanowicz RJ, Meeker M, La Cava W, Olson RS, Moore JH (2018) Relief-based feature selection: Introduction and review. J Biomed Inform 85:189–203
70. Jia J, Yang N, Zhang C, Yue A, Yang J, Zhu D (2013) Object-oriented feature selection of high spatial resolution images using an improved relief algorithm. Math Comput Model 58(3):619–626. Computer and Computing Technologies in Agriculture 2011 and Computer and Computing Technologies in Agriculture 2012

71. Xue B, Zhang M, Browne WN, Yao X (2016) A survey on evolutionary computation approaches to feature selection. IEEE Trans Evol Comput 20(4):606–626
72. Palma-Mendoza R-J, de Marcos L, Rodriguez D, Alonso-Betanzos A (2019) Distributed correlation-based feature selection in spark. Inf Sci 496:287–299
73. Chormunge S, Jena S (2018) Correlation based feature selection with clustering for high dimensional data. J Electr Syst Inf Technol 5(3):542–549
74. https://www.yelp.com/dataset, Accessed: 2020-12-22
75. https://github.com/cjhutto/vaderSentiment, Accessed: 2020-12-22
76. http://www.cs.jhu.edu/~mdredze/datasets/sentiment/, Accessed: 2020-12-22
77. http://www.cs.cornell.edu/people/pabo/movie-review-data/, Accessed: 2020-12-22
78. https://alt.qcri.org/semeval2014/task4/, Accessed: 2020-12-22
79. http://www.cad.zju.edu.cn/home/dengcai/Data/TextData.html, Accessed: 2020-12-21
80. http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/, Accessed: 2020-12-21
81. http://qwone.com/~jason/20Newsgroups/, Accessed: 2020-12-21
82. http://gabrilovich.com/resources/data/techtc/techtc100/, Accessed: 2020-12-21
83. Lloret E, Romá-Ferri MT, Palomar M (2013) Compendium: A text summarization system for generating abstracts of research papers. Data Knowl Eng 88:164–175
84. Erkan G, Radev DR (2004) Lexrank: Graph-based lexical centrality as salience in text summarization. J Artif Int Res 22(1):457–479
85. Saranyamol CS, Sindhu L (2014) A Survey on Automatic Text Summarization. Int J Comput Sci Inf Technol 5
86. Gupta V, Lehal GS (2010) A Survey of Text Summarization Extractive Techniques. J Emerg Technol Web Intell 2:258–268
87. Akhtar MS, Gupta D, Ekbal A, Bhattacharyya P (2017) Feature selection and ensemble construction: A two-step method for aspect based sentiment analysis. Knowl-Based Syst 125:116–135
88. Hatzivassiloglou V, McKeown KR (1997) Predicting the semantic orientation of adjectives. In: 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, Madrid, pp 174–181
89. Miller GA (1995) Wordnet: A lexical database for english. Commun ACM 38(11):39–41
90. Gupta VK, Siddiqui TJ (2012) Multi-document summarization using sentence clustering. In: 2012 4th International Conference on Intelligent Human Computer Interaction (IHCI), pp 1–5
91. El-Kassas WS, Salama CR, Rafea AA, Mohamed HK (2021) Automatic text summarization: A comprehensive survey. Expert Syst Appl 165:113679
92. Toh Z, Wang W (2014) Dlirec: Aspect term extraction and term polarity classification system. In: SemEval@COLING
93. Bharti KK, Singh PK (2016) Opposition chaotic fitness mutation based adaptive inertia weightBPSO for feature selection in text clustering. Appl Soft Comput 43:20–34
94. Baccianella S, Esuli A, Sebastiani F (2013) Using micro-documents for feature selection: The case of ordinal text classification. Expert Syst Appl 40:4687–4696
95. Yousefpour A, Ibrahim R, Hamed HNA (2017) Ordinal-based and Frequency-based Integration of Feature Selection Methods for Sentiment Analysis. Expert Syst Appl 75:80–93
96. Sundararaman A, Ramanathan SV, Thati R (2018) Novel Approach to Predict Hospital Readmissions Using Feature Selection from Unstructured Data with Class Imbalance. Big Data Res 13:65–75
97. Manochandar S., Punniyamoorthy M. (2018) Scaling Feature Selection Method for Enhancing the Classification Performance of Support Vector Machines in Text Mining. Comput Ind Eng 124:139–156
98. Al-Salemi B, Ayob M, Noah SAM (2018) Feature ranking for enhancing boosting-based multi-label text categorization. Expert Syst Appl 113:531–543
99. Yarlagadda M, Rao KG, Srikrishna A (2019) Frequent itemset-based feature selection and Rider Moth Search Algorithm for document clustering. Journal of King Saud University - Computer and Information Sciences
100. Rehman A, Javed K, Babri HA (2017) Feature selection based on a normalized difference measure for text classification. Inf Process Manag 53:473–489
101. Zhang Y, Wanga S, Phillips P, Ji G (2014) Binary PSO with mutation operator for feature selection using decision tree applied to spam detection. Knowl-Based Syst 64:22–31
102. Sanghani G, Kotecha K (2019) Incremental Personalized E-mail Spam Filter using Novel TFDCR Feature Selection with Dynamic Feature Update. Expert Syst Appl 115:287–299
103. Metin SK (2018) Feature selection in multiword expression recognition. Expert Syst Appl 92:106–123
104. Karabulut M (2013) Fuzzy unordered rule induction algorithm in text categorization on top of geometric particle swarm optimization term selection. Knowl-Based Syst 54:288–297
105. Lu Y, Liang M, Ye Z, Cao L (2015) Improved particle swarm optimization algorithm and its application in text feature selection. Appl Soft Comput 35:629–636
106. Singh S, Singh AK (2018) Web-spam features selection using cfs-pso. Proced Comput Sci 125:568–575. The 6th International Conference on Smart Computing and Communications
107. Bharti KK, Singh PK (2016) Opposition chaotic fitness mutation based adaptive inertia weight bpso for feature selection in text clustering. Appl Soft Comput 43:20–34
108. Chen L, Jiang L, Li C (2021) Modified dfs-based term weighting scheme for text classification. Expert Syst Appl 168:114438
109. Liu Y, Loh HT, Sun A (2009) Imbalanced text classification: A term weighting approach. Expert Syst Appl 36(1):690–701
110. Lan M, Tan CL, Su J, Lu Y (2009) Supervised and traditional term weighting methods for automatic text categorization. IEEE Trans Pattern Anal Mach Intell 31(4):721–735
111. Dogan T, Uysal AK (2019) Improved inverse gravity moment term weighting for text classification. Expert Syst Appl 130:45–59
112. Uguz H (2011) A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. Knowl-Based Syst 24(7):1024–1032
113. Shang C, Li M, Feng S, Jiang Q, Fan J (2013) Feature selection via maximizing global information gain for text classification. Knowl-Based Syst 54:298–309
114. Sadeghian Z, Akbari E, Nematzadeh H (2021) A hybrid feature selection method based on information theory and binary butterfly optimization algorithm. Eng Appl Artif Intell 97:104079
115. Lee C, Lee GG (2006) Information gain and divergence-based feature selection for machine learning-based text categorization. Inf Process Manag 42(1):155–165. Formal Methods for Information Retrieval
116. Wang D, Zhang H, Liu R, Lv W (2012) Feature selection based on term frequency and t-test for text categorization. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, CIKM '12. Association for Computing Machinery, New York, pp 1482–1486

117. Wang D, Zhang H, Liu R, Lv W, Wang D (2014) t-test feature selection approach based on term frequency for text categorization. Pattern Recogn Lett 45:1–10
118. Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res 3(null):993–1022
119. Zhou N, Wang L (2007) A modified T-test feature selection method and its application on the HapMap genotype data. Genom Proteom Bioinform:242–249
120. https://www.nist.gov/itl/products-and-services/color-feret-database, Accessed: 2020-12-21
121. http://vision.ucsd.edu/~leekc/ExtYaleDatabase/ExtYaleB.html, Accessed: 2020-12-21
122. http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html, Accessed: 2020-12-21
123. http://www.cbsr.ia.ac.cn/IrisDatabase.htm, Accessed: 2020-12-21
124. http://vis-www.cs.umass.edu/lfw/, Accessed: 2020-12-21
125. Tang K, Hou X, Shao Z, Ma L (2017) Deep feature selection and projection for cross-age face retrieval. In: 2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), pp 1–7
126. Cheng F, Yu J, Xiong H (2010) Facial expression recognition in jaffe dataset based on gaussian process classification. IEEE Trans Neural Netw 21(10):1685–1690
127. http://www.ehu.eus/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes, Accessed: 2020-12-21
128. http://engineering.purdue.edu/biehl/MultiSpec, Accessed: 2020-12-21
129. http://www.csr.utexas.edu/hyperspectral, Accessed: 2020-12-21
130. http://cancerimagingarchive.net/, Accessed: 2020-12-21
131. http://leo.ugr.es/elvira/DBCRepository/, Accessed: 2020-12-21
132. https://www.ncbi.nlm.nih.gov/, Accessed: 2020-12-21
133. https://luna16.grand-challenge.org/Data/, Accessed: 2020-12-21
134. https://www.kaggle.com/navoneel/brain-mri-images-for-brain-tumor-detection/metadata, Accessed: 2020-12-21
135. Dhal P, Azad C (2020) A novel approach for blood vessel segmentation with exudate detection in diabetic retinopathy. In: 2020 International Conference on Artificial Intelligence and Signal Processing (AISP), pp 1–6
136. Dabbaghchian S, Aghagolzadeh A, Moin MS (2007) Feature extraction using discrete cosine transform for face recognition. In: 2007 9th International Symposium on Signal Processing and Its Applications, pp 1–4
137. Wang X, Ding X, Liu C (2002) Optimized gabor filter based feature extraction for character recognition. In: Object recognition supported by user interaction for service robots, vol 4, 223–226
138. Arivazhagan S, Ganesan L, Angayarkanni V (2005) Color texture classification using wavelet transform. In: Sixth International Conference on Computational Intelligence and Multimedia Applications (ICCIMA'05), pp 315–320
139. Lowe DG (1999) Object recognition from local scale-invariant features. In: Proceedings of the Seventh IEEE International Conference on Computer Vision, vol 2, pp 1150–1157
140. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol 1, pp 886–893
141. Calonder M, Lepetit V, Strecha C, Fua P (2010) Brief: Binary robust independent elementary features. In: Daniilidis K, Maragos P, Paragios N (eds) Computer Vision – ECCV 2010. Springer, Berlin, pp 778–792
142. Kumar RM (2014) A survey on image feature descriptors
143. Bober M (2001) Mpeg-7 visual shape descriptors. IEEE Trans Circ Syst Video Technol 11(6):716–719
144. Zhang S, Huang J, Huang Y, Yu Y, Li H, Metaxas DN (2010) Automatic image annotation using group sparsity. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp 3312–3319
145. Lacombe T, Favreliere H, Pillet M (2020) Modal features for image texture classification. Pattern Recogn Lett 135:249–255
146. Tahir F, Fahiem MA (2014) A Statistical-Textural-Features Based Approach for Classification of Solid Drugs Using Surface Microscopic Images. Computational and Mathematical Methods in Medicine, Hindawi Publishing Corporation 2014
147. Ojala T, Pietikäinen M, Harwood D (1996) A comparative study of texture measures with classification based on featured distributions. Pattern Recogn 29(1):51–59
148. Leandro D, Vignolo DHM (2013) Feature selection for face recognition based on multi-objective evolutionary wrappers. Expert Syst Appl 40:5077–5084
149. Sharma V., Juglan K. C. (2018) Automated Classification of Fatty and Normal Liver Ultrasound Images Based on Mutual Information Feature Selection. IRBM 39:313–323
150. Wang Y-Y, Peng W-X, Qiu C-H, Jiang J, Xia S-R (2019) Fractional-order Darwinian PSO-Based Feature Selection for Media-Adventitia Border Detection in Intravascular Ultrasound Images. Ultrasonics 92:1–7
151. Yurtkan K, Demırel H (2014) Feature selection for improved 3D facial expression recognition. Pattern Recogn Lett 38:26–33
152. Ajit Krisshnaa N. L., Kadetotad Deepak V., Manikantan K., Ramachandran S. (2014) Face recognition using transform domain feature extraction and PSO-based feature selection. Appl Soft Comput 22:141–161
153. Mistry KK, Zhang L, Neoh SC, Lim CP, Fielding B (2016) A Micro-GA Embedded PSO Feature Selection Approach to Intelligent Facial Emotion Recognition. IEEE Trans Cybern 47:1496–1509
154. Peralta D, Triguero I, García S, Saeys Y, Benitez JM, Herrera F (2017) Distributed incremental fingerprint identification with re duce d database penetration rate using a hierarchical classification based on feature fusion and selection. Knowl-Based Syst 126:91–103
155. Jiang Y, Li C (2015) mrmr-based feature selection for classification of cotton foreign matter using hyperspectral imaging. Comput Electron Agric 119:191–200
156. Fu Y, Jia X, Huang W, Wang J (2014) A comparative analysis of mutual information based feature selection for hyperspectral image classification. In: 2014 IEEE China Summit International Conference on Signal and Information Processing (ChinaSIP), pp 148–152
157. Ajit Krisshna NL, Deepak VK, Manikantan K, Ramachandran S (2014) Face recognition using transform domain feature extraction and pso-based feature selection. Appl Soft Comput 22:141–161
158. Shetty S, Kelkar P, Manikantan K, Ramachandran S (2013) Shift invariance based feature extraction and weighted bpso based feature selection for enhanced face recognition. Procedia Technol 10:822–830. First International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA) 2013
159. Aneesh MU, Masand AAK, Manikantan K (2012) Optimal feature selection based on image pre-processing using accelerated binary particle swarm optimization for enhanced face recognition. Procedia Eng 30:750–758. International Conference on Communication Technology and System Design 2011
160. López-Franco C, Villavicencio L, Arana-Daniel N, Alanis AY (2014) Image Classification Using PSO-SVM and an RGB-D Sensor. Mathematical Problems in Engineering, Hindawi 2014

161. Li B, Lai Y-K, Rosin PL (2017) Example-based image colorization via automatic feature selection and fusion. Neurocomputing 266:687–698
162. Li J, Wang JZ, Wiederhold G (2000) Classification of textured and non-textured images using region segmentation. In: Proceedings 2000 International Conference on Image Processing (Cat. No.00CH37101), vol 3, pp 754–757
163. Busso C, Bulut M, Lee C-C, Kazemzadeh A, Mower E, Kim S, Chang JN, Lee S, Narayanan SS (2008) IEMOCAP: interactive emotional dyadic motion capture database. Lang Resour Eval 42(4):335–359
164. Burkhardt F, Paeschke A, Rolfes M, Sendlmeier WF, Weiss B (2005) A database of german emotional speech. In: INTERSPEECH 2005 - eurospeech, 9th european conference on speech communication and technology, ISCA, Lisbon, pp 1517–1520
165. http://kahlan.eps.surrey.ac.uk/savee/Download.html, Accessed: 2020-12-21
166. Tao J, Liu F, Zhang M, Jia H (2008) Design of speech corpus for mandarin text to speech
167. Martin O, Kotsia I, Macq B, Pitas I (2006) The enterface' 05 audio-visual emotion database. In: 22nd International Conference on Data Engineering Workshops (ICDEW'06), pp 8–8
168. G. C, I. I, A. P, M. T (2014) Emovo corpus: An italian emotional speech database, pp 3501–3504
169. Wu T, Yang Y, Wu Z, Li D (2006) Masc: A speech corpus in mandarin for emotion analysis and affective speaker recognition. In: 2006 IEEE Odyssey - The Speaker and Language Recognition Workshop, pp 1–5
170. Akçay MB, Oğuz K (2020) Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Comm 116:56–76
171. Yasmin G, Das AK, Nayak J, Pelusi D, Ding W (2020) Graph based feature selection investigating boundary region of rough set for language identification. Expert Syst Appl 158:113575
172. Bóna J (2014) Temporal characteristics of speech: The effect of age and speech style. J Acoust Soc Amer 136(2):EL116–EL121
173. Caka N (2015) What are the spectral and temporal features in speech signal?
174. Teixeira JP, Oliveira C, Lopes C (2013) Vocal acoustic analysis - jitter, shimmer and hnr parameters. Procedia Technol 9:1112–1122. CENTERIS 2013 - Conference on ENTERprise Information Systems / ProjMAN 2013 - International Conference on Project MANagement/ HCIST 2013 - International Conference on Health and Social Care Information Systems and Technologies
175. Özseven T (2019) A novel feature selection method for speech emotion recognition. Appl Acoust 146:320–326
176. Casale S, Russo A, Serrano S (2007) Multistyle classification of speech under stress using feature subset selection based on genetic algorithms. Speech Commun 49:801–810
177. Mencattini A, Martinelli E, Costantini G, Todisco M, Basile B, Bozzali M, Di Natal C (2014) Speech emotion recognition using amplitude modulation parameters and a combined feature selection procedure. Knowl-Based Syst 63:68–81
178. Rong J, Li G, Chen Y-PP (2009) Acoustic feature selection for automatic emotion recognition from speech. Inf Process Manag 45:315–328
179. Liu Z-T, Wu M, Cao W-H, Mao J-W, Xu J-P, Tan G-Z (2018) Speech emotion recognition based on feature selection and extreme learning machine decision tree. Neurocomputing 273:271–280

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Pradip Dhal is currently pursuing his Ph.D. degree at the National Institute of Technology, Jamshedpur, India. He received his M.Tech degree from the Central University of South Bihar, Bihar, India, in Computer Science. His research interests include Machine Learning, Pattern Recognition, Data Mining, Image Analysis, Natural language Processing.

Chandrashekhar Azad is currently an Assistant Professor of the Department of Computer Applications, National Institute of Technology, Jamshedpur, India. He received the Ph.D. degree from the Department of Computer Science and Engineering, Birla Institute of Technology, Mesra, Ranchi, Jharkhand, India. His research interests are Data Mining, Machine Learning Intrusion Detection System Medical Mining, and Artificial Intelligence.