# Chapter – 5

# Feature Selection through Clustering to Classify High Dimensional Data

## *5.1 Introduction*

Innovative information technologies are omnipresent in this era. Data is accumulating with enormous speed and in a very inexpensive way. The human mind is incapable of processing it with its rate of growth. Extracting and analyzing useful information from such a huge volume of data needs some automatic statistical and machine learning techniques. Huge sample size with high dimensional datasets suffers from problems such as bogus correlations, and heavy computation cost while a small sample size with high-dimensional data cannot lead to valid conclusions. Eliminating irrelevant features is the efficient way because they exert an undue load on classifiers and increase the time and resources needed to build the model (Ozcift,2011). Moreover, high dimensional data sets may contain groups of correlated attributes they measure the same underlying meaning. The irrelevant dataset can also mislead the logic of the algorithm that affects the performance of the model (Qinbao et.al.,2013).

High dimensional data generally degrades the performance of Data Mining algorithms. To improve the performance, we can reduce dimensionality by eliminating irrelevant, redundant, and uninformative features, as such features only exert undue burden on algorithms and increase the time and resources needed to build the model. The elimination of such features is proposed by the clustering-based feature selection method. The main purpose of this method is to select the most influencing relevant feature space by discarding irrelevant features that lead to more accurate results. The Comparative Analysis of the Proposed Approach on 9 datasets (features ranging from 12 to 2,002) with a different number of features is presented. The efficacy of the proposed model is

compared with the standard correlation-based Feature Selection approach (RELIEF and Info-Gain Approach) on the SVM classifier.

A better feature set can solve numerous machine-learning problems (Butterworth et al. ,2005). Hence Feature selection & Extraction is one of the major fields in Data mining. Two commonly used approaches reduce the dimensionality of the feature set (Fodor,2002).

1) The first one – Transformation of existing features into a less dimensional space by Feature Extraction.

2) The second one – Selection of a subset of original feature space without their transformation (Dew,2016).

Feature selection /extraction is a proven pre-processing step that reduces dimensionality, increases comprehensibility, improves the overall performance, and reduces the complexity and computation time of any classification algorithms (Lei & Liu,2003).

Co-relation based clustering algorithms are not only applicable to cluster data based on their similarity but also useful for feature compression, and reduction technique. The correlation-based feature selection method K- means have been proposed to discover the similarity and dissimilarity among data and to form the clusters.

In this study, the high dimensionality problem is discussed and a new clustering-based feature selection method is proposed. The K-means algorithm is used to form clusters based on the similarities of features. Cluster representatives from the set of features are sorted out from each Cluster. Cluster centroid, ranking algorithms, and random selection are the three methods discussed in the literature used to select cluster representatives. The Proposed model used centroid as a cluster representative. The reduced set of features is applied to the SVM classifier and the results are compared with

Relief and IG (Information gain) feature selection methods. Relevant features were tested for the accuracy of the proposed model.

The purpose of feature subset selection is to select and remove as many redundant and irrelevant features as possible. The elimination of such features that are not participating in making predictions improves the performance of all machine-learning algorithms.

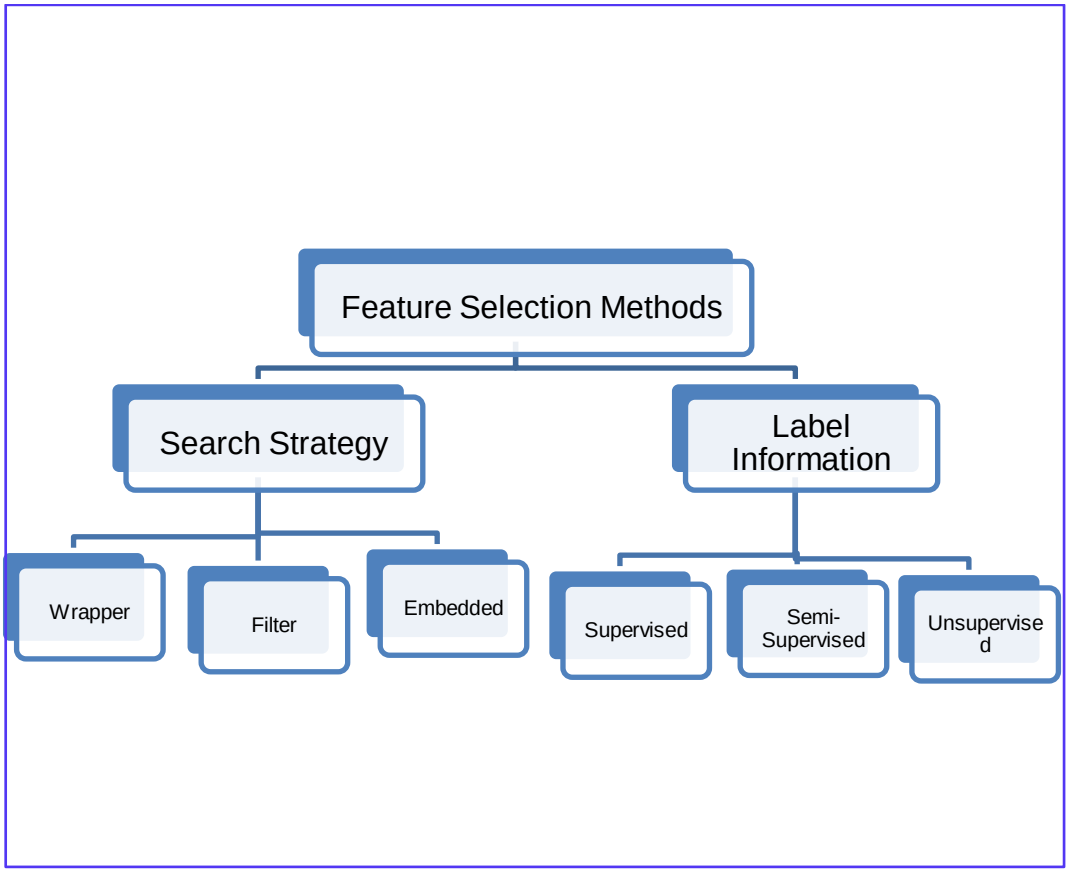



Figure 5.1 Feature Selection Methods

Figure 5.1 is a tree diagram for feature selection methods based on the search strategy used and labeled information provided. According to the provided labeled information, the feature selection technique is classified into three categories: 1. Supervised methods 2. Semi-supervised methods and 3. Unsupervised methods.

The supervised feature selection algorithms can effectively select and discriminate relevant features from different classes when the samples are fully labeled. However, all supervised learning methods have the limitations of the requirement for fully labeled data, which are expensive practice. The performances of supervised learning methods drop dramatically when sufficient labeled data is not provided.

When a mixture of labeled and unlabeled data is given, semi-supervised feature selection can be used to construct a similarity matrix to identify relevant features. When data labels are absent, unsupervised feature selection can be considered to identify the relevant features.

Based on the searching strategies, feature selection techniques can be classified into three categories. The wrapper, Filter, and Embedded Methods. Wrapped is a time-consuming method, it takes many implementations of feature learning algorithms to identify the final feature set. The incapability to handle large and high dimensional datasets limits its applicability.

The filter method identifies the final feature set by understanding the intrinsic properties present in the data. It is a two-step process at the first step rank of the features is given according to criteria. In the second step, highest-ranked features are selected. Relief, F-statistic, and Information gain are few well-known algorithms for feature selection.

In Embedded models feature selection is done in model construction itself.

## *5.2 Introduction about Data set*

Data set from the UC Irvine Machine Learning Repository viz. Heart Disease, Wine quality white, Parkinson dataset, Colon Cancer, Breast Cancer, Image Segmentation, Cleveland-0, Ionosphere, and Squash Harvest Stored are used for the study. Table- 5.1 gives a piece of detailed information about several features, number of instances, and the number of output classes with their name about the datasets used for the study.

Table 5.1 Data set description

| S. No | Dataset Name | #Features | #Instances | Class |
|---|---|---|---|---|
| 1 | Heart Disease | 14 | 303 | 2{Yes, No} |
| 2 | Wine quality white | 12 | 1482 | 2{Negative, Positive} |
| 3 | Parkinson dataset | 756 | 757 | 2{Yes, No} |
| 4 | Colon Cancer | 2002 | 62 | 2{Normal, Abnormal} |
| 5 | Breast Cancer | 32 | 569 | 2{Malignant, Benign} |
| 6 | Image Segmentation | 19 | 210 | 7{ Brickface, Sky, Foliage, Cement, Window, Path, Grass} |
| 7 | Cleveland-0 | 13 | 173 | 2{Negative, Positive} |
| 8 | Ionosphere | 35 | 351 | 2{Good, Bad} |
| 9 | Squash Harvest Stored | 25 | 53 | 3{Excellent, Ok, Not acceptable} |

## *5.3 Preliminaries and basic definitions*

### 5.3.1 SVM Classifier

The support vector machine (SVM) is the most appropriate, effective, and popular non-linear statistical learning method among all classification algorithms. It is especially applicable for diagnosis and prognosis classification problems because of its high generalization capacity (Jakkula,2006).

### 5.3.2 RELIEF Feature Selection Approach

Relief feature selection algorithm proposed by Kira and Rendell (1992) is considered as a simple and efficient method for numerical and nominal attribute selectors (Ryan,2018). The Relief algorithm calculates the weight of each feature depending on its relevance for the class. Initially, weights are equal to zero and iteratively update within the range of -1 to +1. At last, the required numbers of features are selected with the highest positive weights.

### 5.3.3 Info-Gain Feature Selection Approach

Information gain measures the amount of information present in its bit to make predictions (Pratiwi et al., 2018). Information gain is one of the most t popular feature selection methods. It was presented by Claude Shannon in 1948. Information Gain is used to selecting important features with respect to class attributes and the features having a certain threshold are selected (Verdu, 1998).

## *5.5 Methodology*

The Model involves four steps: data pre-processing, feature extraction /selection to identify and remove irrelevant, redundant, or noisy features from the provided dataset, data classification, and performance evaluation. The reduced dimensional feature set is used as input to the classifiers.

Weka and Orange datamining tools are used to implement the algorithm. Weka 3.8 provides tools for data preprocessing, implementation of several Machine Learning algorithms, and visualization so that one can develop machine learning techniques and apply them to real-world data mining problems. Orange is an open-source software package. 3.0 version is used for the implementation.Orange consists of a canvas interface onto which widgets can be placed to create a data analysis workflow. Widgets offer basic functionalities.Feature selection is performed through Widgets provided by orange tool and classification through SVM classifier is done by Weka tool.The K-Means clustering algorithm is applied to cluster feature set based on their similarities. As an output k number of clusters will be constructed. For each dataset 4 values of K will be examined for example in Heart disease dataset K=3(20%), K=6(40%) ,K=8(60%) and K=11(80%) clusters are formed. 10 fold cross validation method is used for the training and testing of SVM classifier. For all values of K the performance is evaluated on various parameters for proposed model, Relief and Info gain FS methods. Centroid based method is deployed in this implementation.

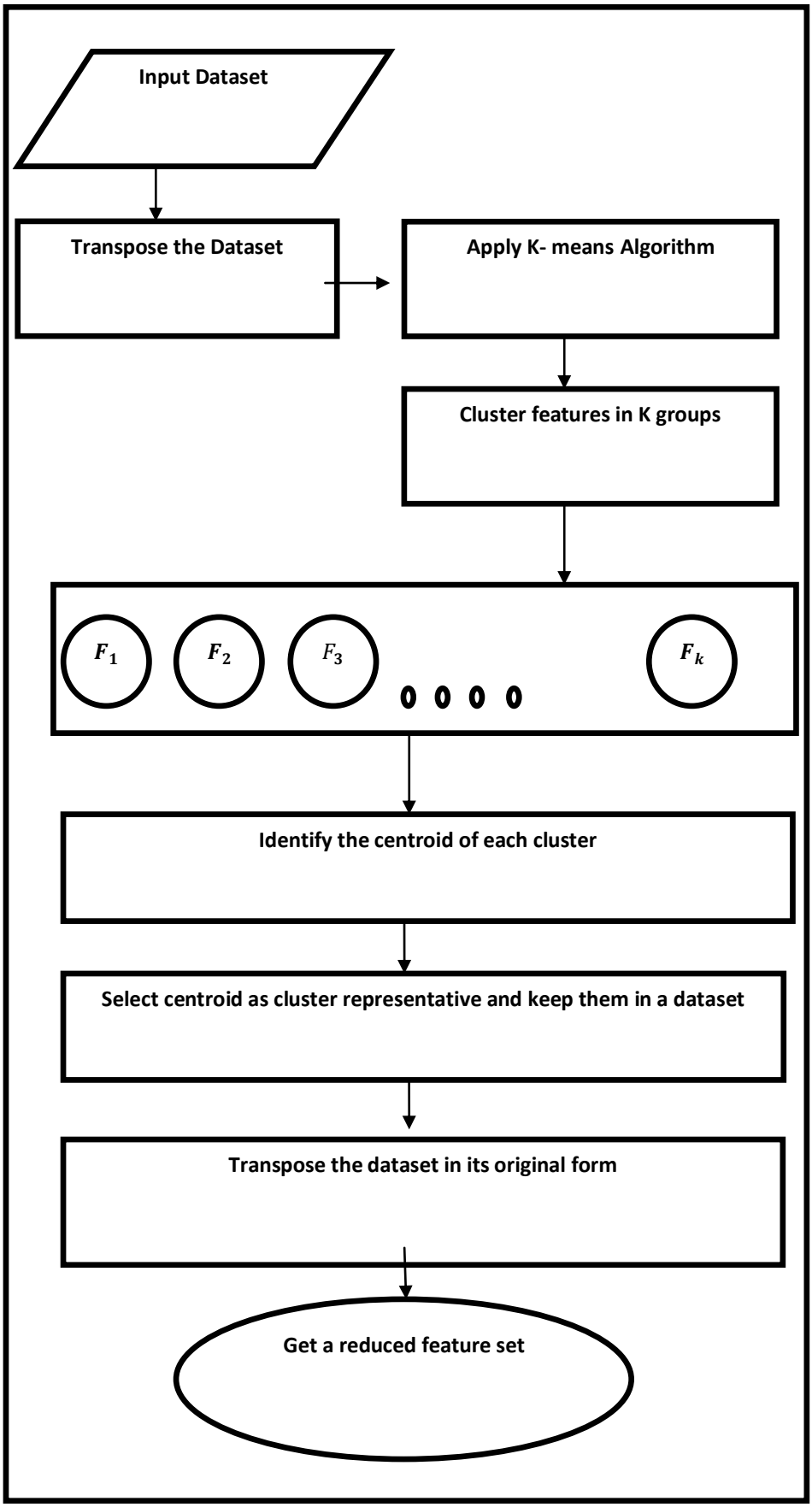



Figure-5.2 Flow of Control for the proposed model

Figure -5.2 is a graphical representation of the model. Data input in the system it is then transformed into its transpose so that the instances/records become columns and features/attributes will become rows. The K-means clustering algorithm will be applied to the cluster feature set based on their similarities. As an output K number of clusters will be constructed.

The advantage of using the clustering algorithm in the study is twofold: at first, all the features having similar characteristics are grouped into one cluster and the then the second would be to conveniently to select few or required features from each cluster, which reflect, nature of the overall sample population. Centroid based, Rank-based, and random selection methods for the selection of cluster representatives is discussed in the literature out of which Centroid based method is deployed.

Algorithm for the proposed model

**Input**: A High Dimensional Data set
N: Number of Clusters (Number of features to be selected)
**Output:** The performance parameters of SVM classifier on a reduced dataset
Procedure:
Step 1: Input High dimensional dataset.
Step 2: Find the Transpose (Features as Rows and Instances as Columns).
Step 3: Input the number of clusters to be constructed.
Step 4: Apply the K-Means algorithm with Euclidean distance.
Step 5: Get the N clusters as the result of K- means algorithm
Step 6: Calculate the centroid of each Cluster
Step 7: Get the centroid from each Cluster and keep them in a separate dataset
Step 7: Transpose the dataset
Step 8: Apply the SVM classifier and observe the performance of the parameters.

Table 5.2 Comparative analysis of the Proposed Approach with Relief & Info-Gain Feature Selection approach

| Dataset | % of feature selected | No of features selected | Proposed Approach: (Clustering-based Feature Selection Model) | | | RELIEF: Feature Selection Approach | | | Info-Gain: Feature Selection Approach | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | F-Measure | ROC Area | Accuracy | F-Measure | ROC Area | Accuracy | F-Measure | ROC Area | Accuracy |
| Heart Disease Dataset | 20 % | 3 | .73 | .69 | 70.0 | .79 | .77 | 77.5 | .83 | .80 | 81.1 |
| | 40 % | 6 | .79 | .76 | 76.8 | .81 | .78 | 78.8 | .85 | .82 | 83.1 |
| | 60 % | 8 | 0.96 | 0.96 | 96.3 | .84 | .81 | 82.1 | .85 | .82 | 83.1 |
| | 80 % | 11 | .81 | .78 | 79.2 | .85 | .82 | 82.8 | .85 | .82 | 83.4 |
| | Whole Data set | 14 | .85 | .82 | 83.4 | .85 | .82 | 83.4 | .85 | .82 | 83.4 |
| Wine-quality white Dataset | 20 % | 3 | .99 | .98 | 98.3 | .97 | .97 | 97.8 | .99 | .50 | 98.3 |
| | 40 % | 5 | .99 | .98 | 98.3 | .99 | .98 | 98.3 | .99 | .50 | 98.3 |
| | 60 % | 8 | .99 | .98 | 98.3 | .99 | .98 | 98.3 | .99 | .50 | 98.3 |
| | 80 % | 10 | .99 | .98 | 98.3 | .99 | .98 | 98.3 | .99 | .50 | 98.3 |
| | Whole Data set | 12 | .99 | .98 | 98.3 | .99 | .98 | 98.3 | .99 | .98 | 98.3 |
| Parkinson Dataset | 20 % | 151 | .88 | .74 | 85.0 | .88 | .79 | 85.5 | .88 | .76 | 86.3 |
| | 40 % | 302 | .89 | .74 | 86.2 | .89 | .79 | 86.1 | .89 | .78 | 86.1 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 60 % | 454 | .92 | .78 | 86.4 | .91 | .78 | 86.1 | .91 | .78 | 86.0 |
| | 80 % | 605 | .90 | .78 | 85.9 | .90 | .77 | 85.4 | .90 | .78 | 85.4 |
| | Whole Data set | 756 | .90 | .78 | 85.7 | .90 | .78 | 85.7 | .90 | .78 | 85.7 |
| Colon Dataset | 20 % | 400 | .88 | .83 | 85.4 | .84 | .75 | 79.0 | .84 | .75 | 79.0 |
| | 40 % | 800 | .88 | .83 | 85.4 | .85 | .76 | 80.6 | .85 | .76 | 80.6 |
| | 60 % | 1201 | .88 | .83 | 85.4 | .86 | .79 | 82.2 | .85 | .76 | 80.6 |
| | 80 % | 1602 | .88 | .83 | 85.4 | .86 | .79 | 82.2 | .86 | .79 | 82.2 |
| | Whole Data set | 2002 | .88 | .83 | 85.4 | .88 | .83 | 85.4 | .88 | .83 | 85.4 |
| Breast cancer Dataset | 20 % | 7 | .90 | .92 | 93.1 | .94 | .95 | 96.3 | .91 | .93 | 94.0 |
| | 40 % | 13 | .92 | .93 | 94.5 | .95 | .95 | 96.8 | .91 | .92 | 93.6 |
| | 60 % | 19 | .95 | .95 | 97.6 | .96 | .96 | 96.5 | .95 | .95 | 96.6 |
| | 80 % | 26 | .96 | .96 | 97.3 | .97 | .97 | 97.8 | .97 | .97 | 97.8 |
| | Whole Data set | 32 | .97 | .97 | 97.8 | .97 | .97 | 97.8 | .97 | .97 | 97.8 |
| Image segmentation Dataset | 20 % | 4 | .47 | .84 | 57.6 | .28 | .74 | 49.0 | .37 | .76 | 50.0 |
| | 40 % | 8 | .65 | .91 | 75.2 | .48 | .84 | 70.4 | .48 | .89 | 72.3 |
| | 60 % | 11 | .73 | .95 | 95.6 | .72 | .90 | 87.6 | .70 | .89 | 90.0 |
| | 80 % | 15 | .69 | .92 | 85.7 | .73 | .95 | 89.0 | .70 | .89 | 87.6 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Whole Data set | 19 | .80 | .96 | 88.5 | .80 | .96 | 88.5 | .80 | .96 | 88.5 |
| Cleland Dataset | 20 % | 3 | .96 | .50 | 92.4 | .97 | .76 | 94.7 | .96 | .57 | 93.6 |
| | 40 % | 6 | .96 | .68 | 93.6 | .97 | .68 | 94.7 | .97 | .76 | 95.9 |
| | 60 % | 9 | .98 | .80 | 96.5 | .97 | .76 | 95.3 | .97 | .76 | 95.3 |
| | 80 % | 11 | .98 | .80 | 96.5 | .97 | .80 | 95.9 | .97 | .76 | 95.3 |
| | Whole Data set | 14 | .97 | .80 | 95.9 | .97 | .80 | 95.9 | .97 | .80 | 95.9 |
| Ionosphere Dataset | 20 % | 7 | .89 | .82 | 86.0 | .87 | .77 | 82.3 | .87 | .77 | 82.3 |
| | 40 % | 14 | .90 | .83 | 86.6 | .89 | .82 | 86.0 | .89 | .82 | 86.0 |
| | 60 % | 21 | .89 | .82 | 85.7 | .88 | .78 | 83.4 | .88 | .78 | 83.4 |
| | 80 % | 28 | .90 | .82 | 86.3 | .90 | .82 | 86.3 | .88 | .78 | 83.4 |
| | Whole Data set | 35 | .82 | .85 | 88.6 | .82 | .85 | 88.6 | .82 | .85 | 88.6 |
| Squash Dataset | 20 % | 5 | .57 | .61 | 50 | .73 | .74 | 59.6 | .73 | .77 | 59.6 |
| | 40 % | 10 | .65 | .72 | 61.5 | .69 | .74 | 57.6 | .69 | .71 | 57.6 |
| | 60 % | 15 | .73 | .78 | 63.4 | .64 | .71 | 53.8 | .77 | .81 | 67.3 |
| | 80 % | 20 | .73 | .78 | 63.4 | .76 | .84 | 73.0 | .73 | .79 | 67.3 |
| | Whole Data set | 25 | .76 | .83 | 71.1 | .76 | .83 | 71.1 | .76 | .83 | 71.1 |

The proposed model is tested on the different number of selected features with the diversified dataset and the results are presented in Table 5.2. It can be identified that some datasets give good performance with less number of selected features whereas some datasets need approximately all feature for better performance.

## *5.6 Performance analysis of the Generated Model*

The objective of this experiment is to minimize the feature set & to maximize the accuracy of the classifier. Thus, a different percentage of features are selected and tested on some standard performance measuring parameters such as F-Measure, ROC, and Accuracy. The three methods discussed in the literature survey for selection of cluster representatives are Random feature selection, the top-ranked feature, or cluster centroid from each cluster.

The graphical representation of several selected features with the accuracy of the proposed model against Relief and Info-Gain is shown in Figure 5.3 to 5.11. The graphical representation of several selected features with the F-measure of the proposed model against Relief and Info-Gain is also shown in Figures.

The SVM classifier's performance on Heart disease reduced the percentage of datasets through the proposed model, Relief, and Info –gain revealed (figure-5.3) that the Accuracy of the proposed model with 8 features (60%) is highest.

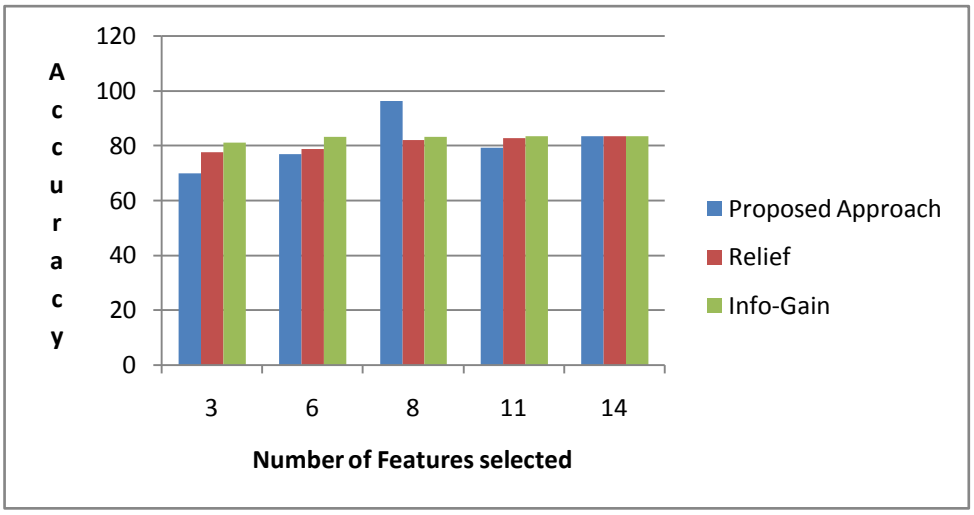



**Figure 5.3 Chart for Heart Disease dataset Accuracy**

Figure- 5.4 displays that the Accuracy of the proposed model is the same as the other two feature selection method in the case of Wine quality white dataset. But the performance of Relief is very poor with fewer features i.e. with 10%.

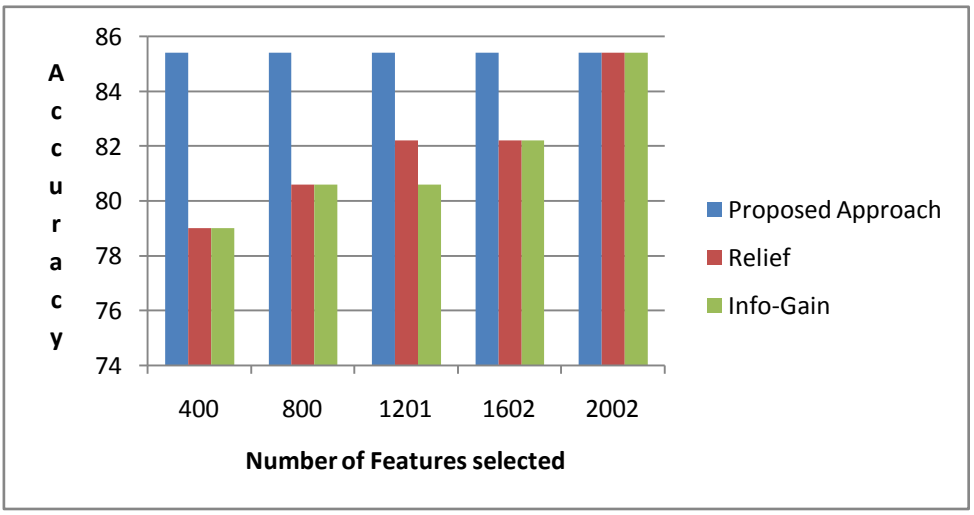



**Figure 5.4: Chart for Wine quality white dataset Accuracy**

With the Parkinson dataset, the accuracy of the proposed model is reported highest with 454(40%) of feature displayed in Figure-5.5.

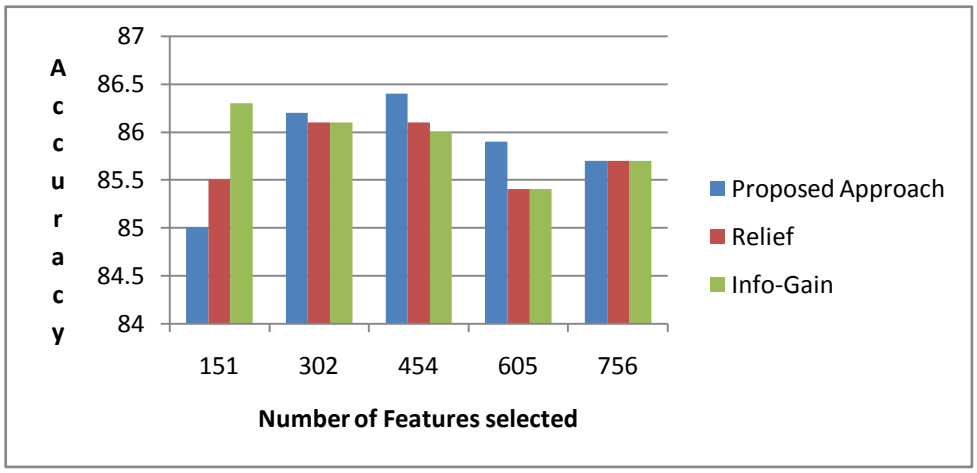



**Figure 5.5: Chart for Parkinson dataset Accuracy**

In the case of the Colon cancer dataset, the accuracy of the proposed model is excellent compare to the other two state-of-art methods, in all percentages of selected features, it is shown in figure-5.6.

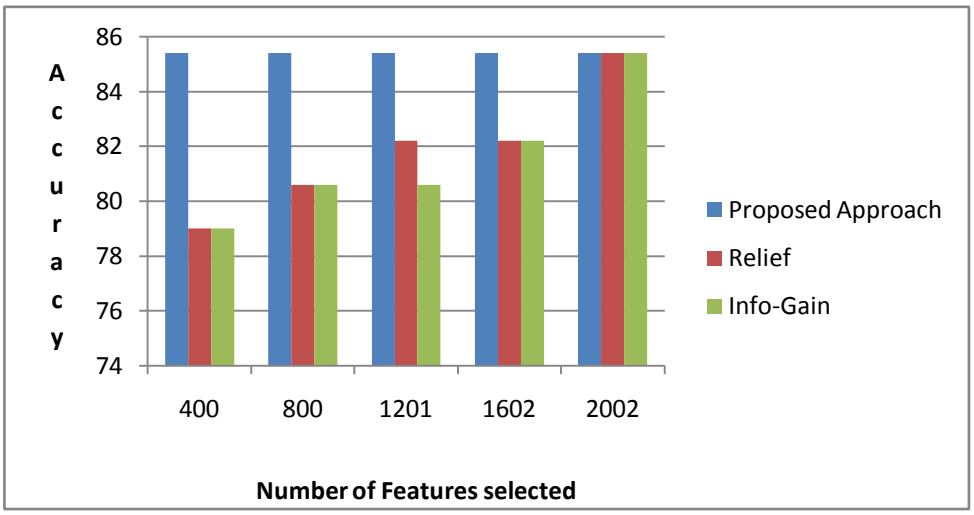



**Figure 5.6: Chart for Colon Cancer dataset Accuracy**

In Figure- 5.7 the performance of the proposed model, Info gain, and relief is compared by the results of SVM classifier on Brest cancer dataset

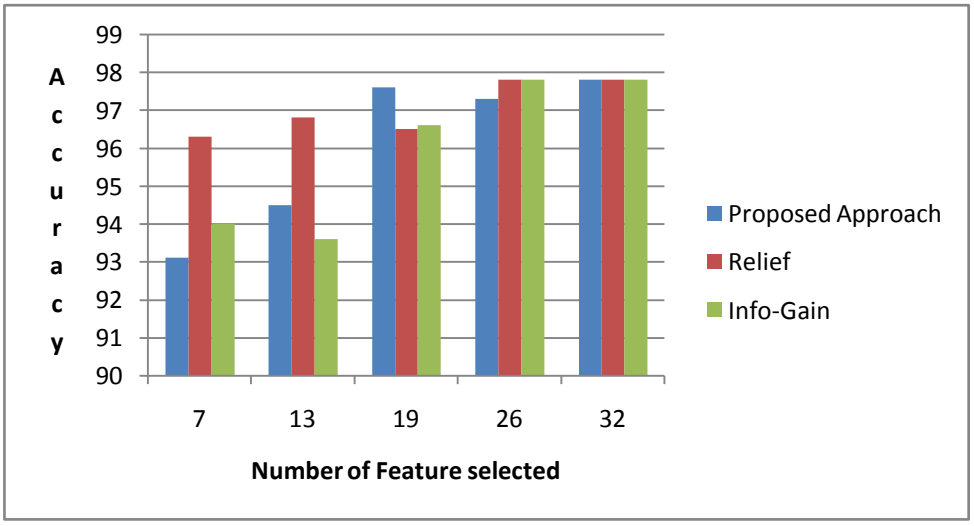



**Figure 5.7: Chart for Breast Cancer dataset Accuracy**

The accuracy of the proposed model is reported high with fewer features i.e. 10%, 20%, and 40 %, and highest with 40 % in Image segmentation dataset figure -5.8.

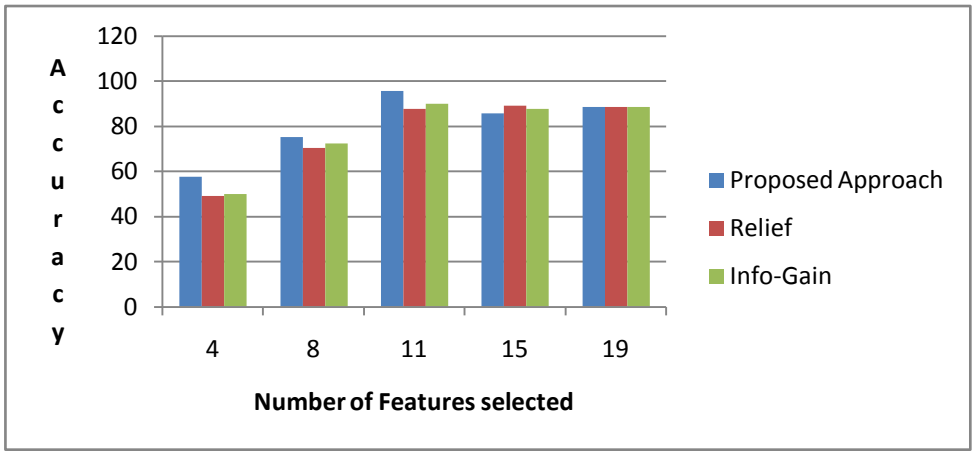



**Figure 5.8: Chart for Image Segmentation dataset Accuracy**

Figure-5.9 is the graph for the Cleland dataset and the accuracy observed is highest with 9(40%) of features using the proposed model.

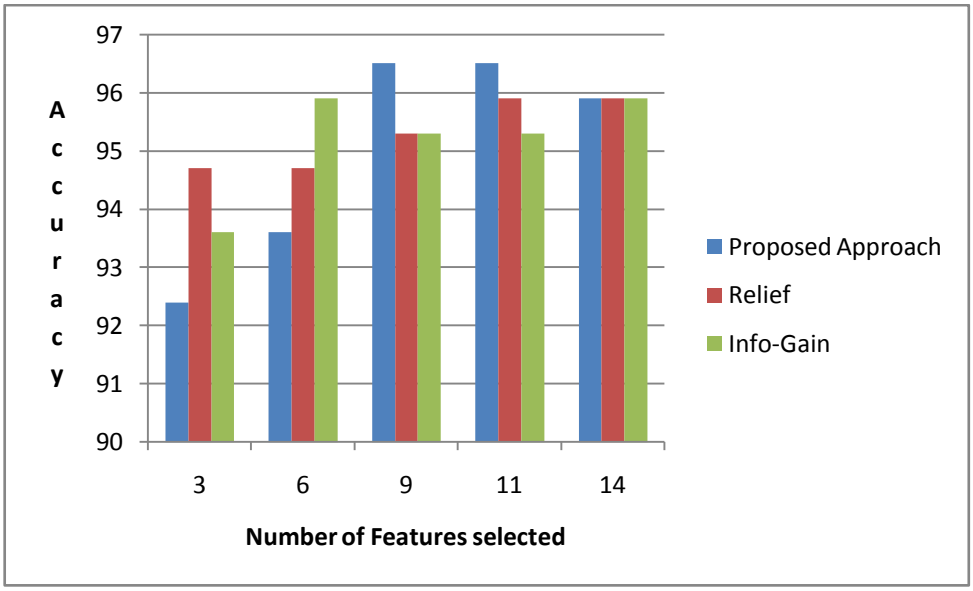



**Figure 5.9: Chart for Cleland dataset Accuracy**

The accuracy of the ionosphere dataset observed highest in all the percentage of features set furthermore it is observed that all the features are essential to get high accuracy (Figure-5.10).

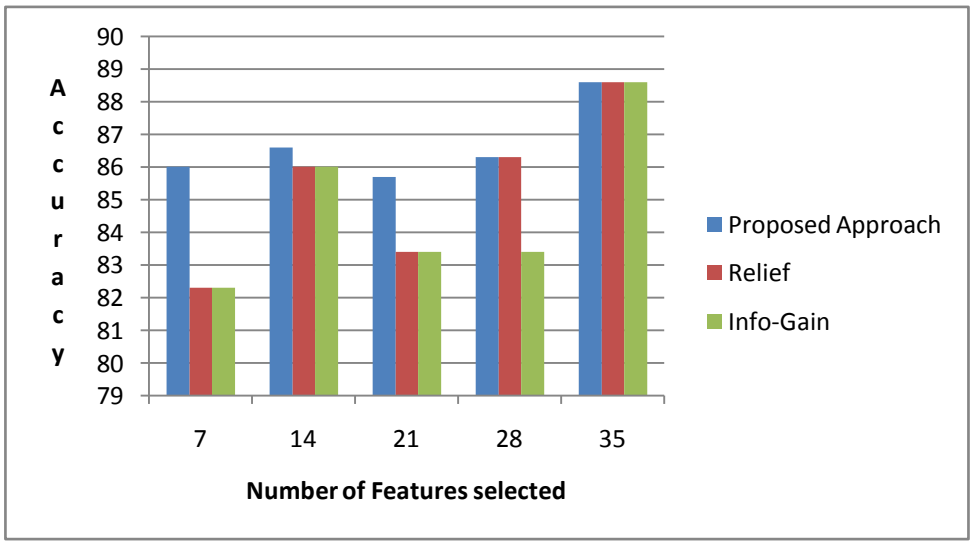



**Figure 5.10: Chart for Ionosphere dataset Accuracy**

In figure -5.11 chart for Squash dataset is given all features are required to achieve high accuracy.

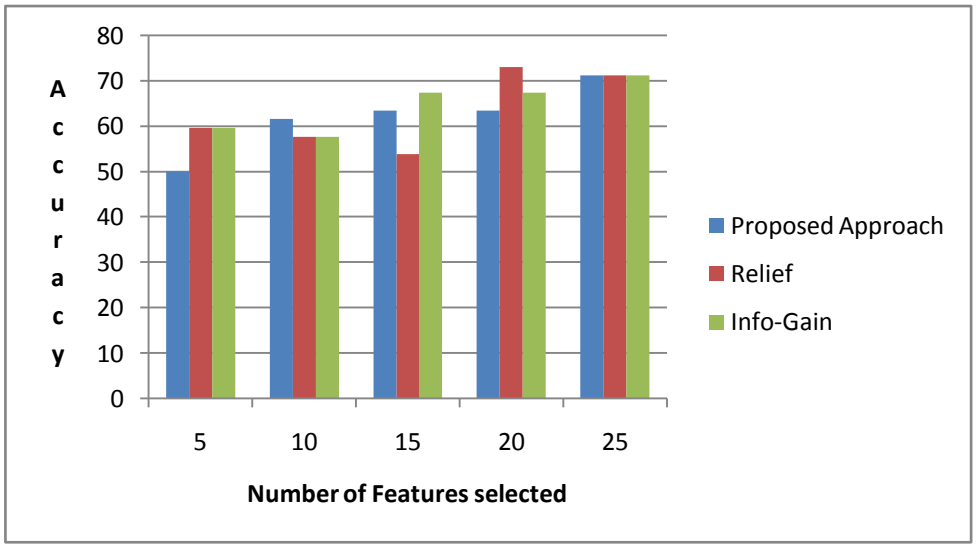
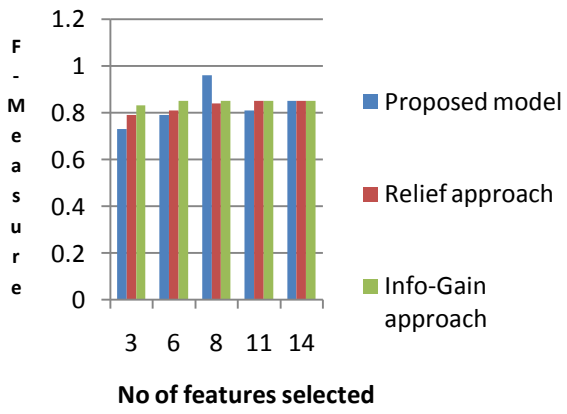


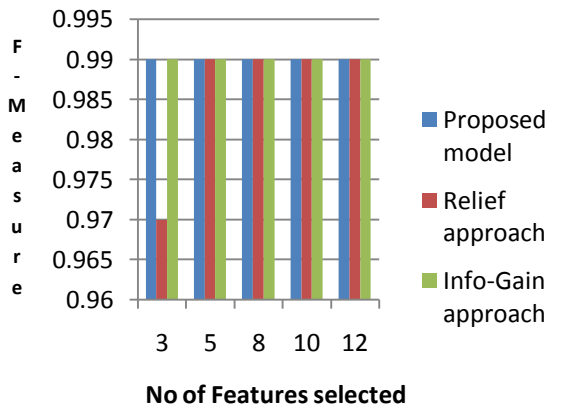

**Figure 5.11: Chart for Squash dataset Accuracy**

The following figures are the charts representing plot on several selected features with F-measure and ROC values. Improved value of F-Measure and ROC with the proposed model is observed.

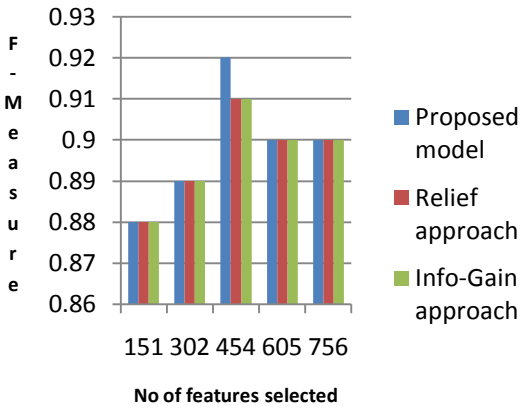
Heart Disease dataset



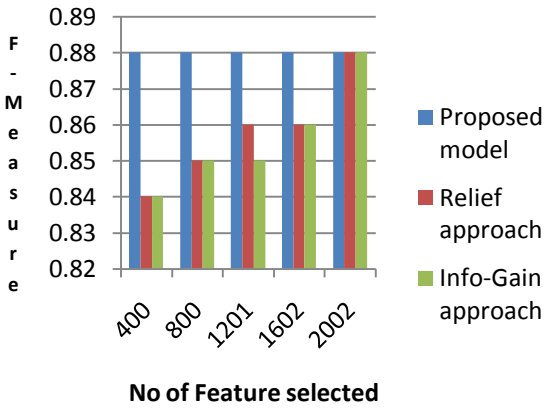

Wine quality White dataset



Parkinson Dataset




Colon dataset




Breast cancer dataset

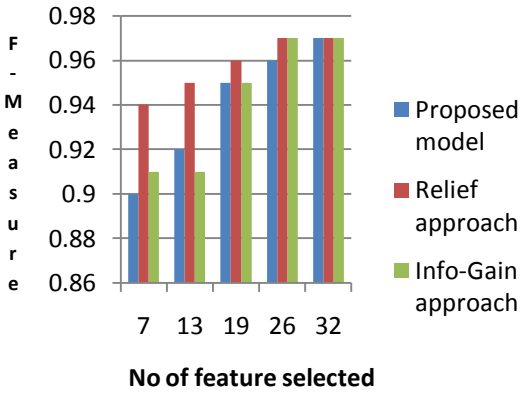
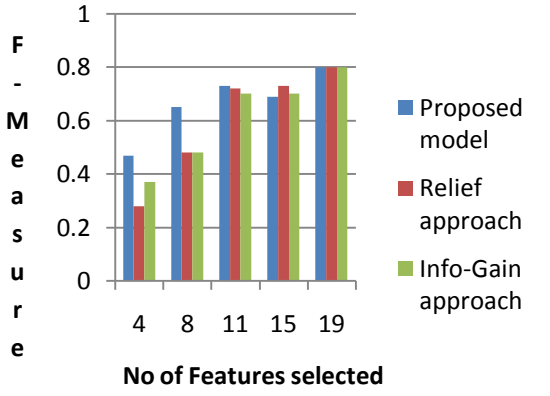
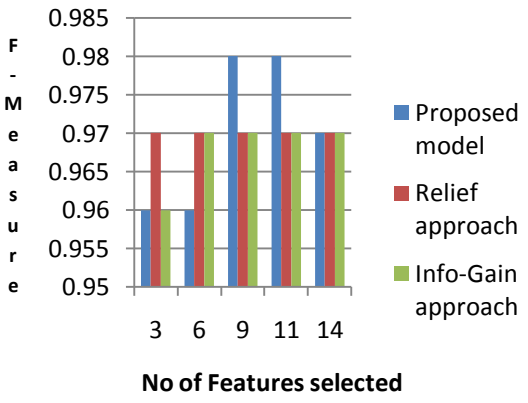


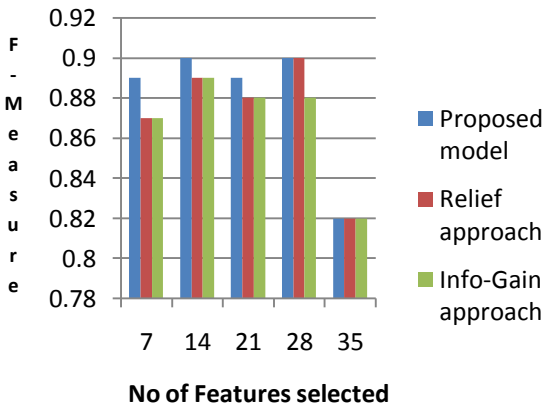

Image segmentation dataset



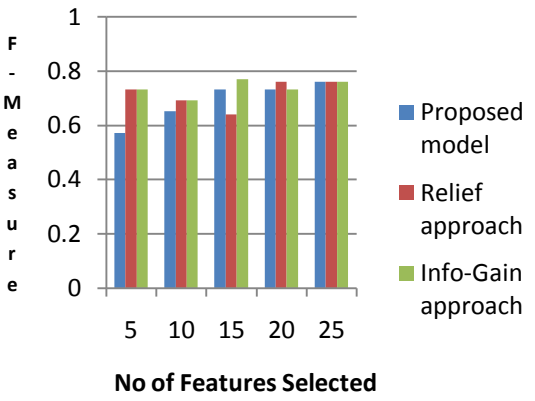

Cleland Dataset



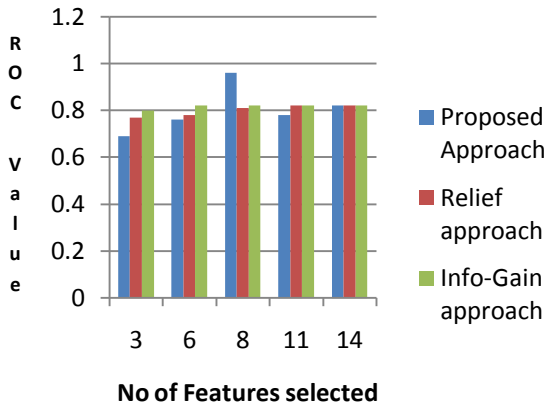

Ionosphere dataset



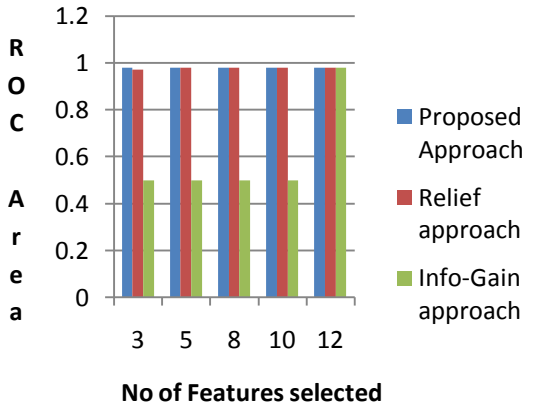
Squash Dataset



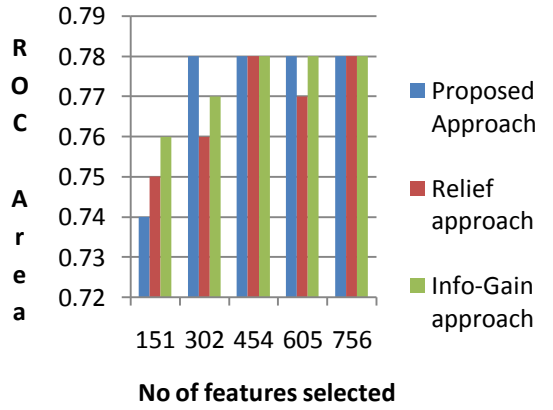

Heart Disease Dataset



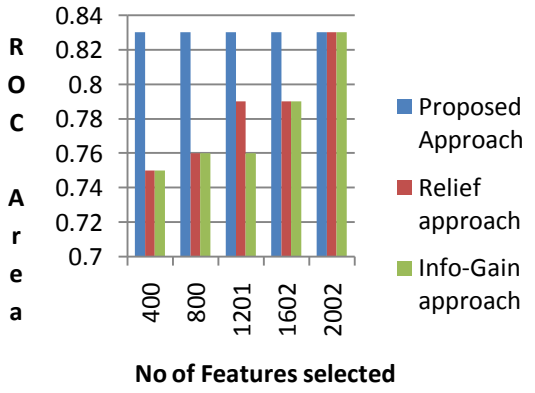

Wine quality White Dataset



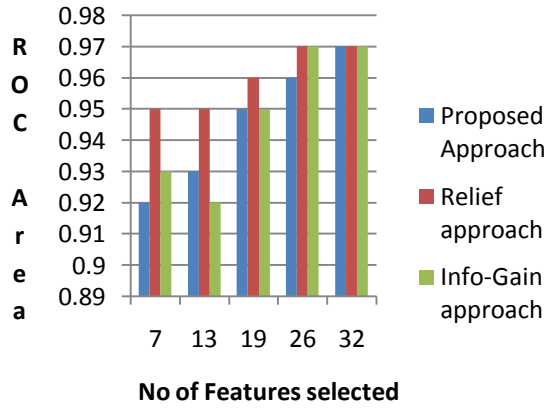

Parkinson Dataset




Colon Dataset




Breast cancer Dataset



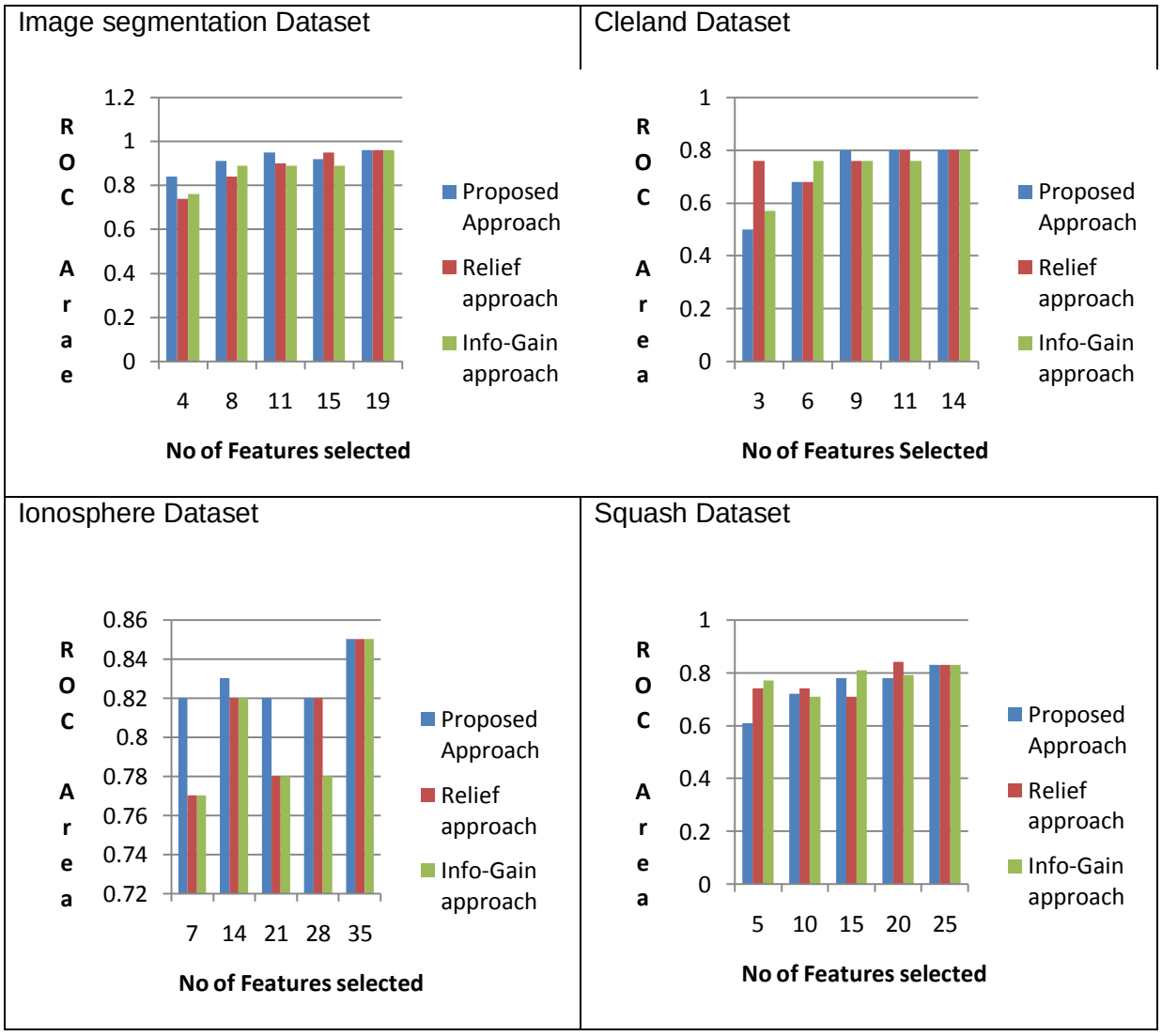



**Figure 5.12: Chart for F-measure and ROC values**

## *5.7 Summary*

The proposed clustering-based feature selection method is to eliminate irrelevant and redundant features from a high dimensional dataset. Table 5.1 gives a comparative analysis of the Proposed Approach with a standard correlation-based Feature Selection approach (RELIEF and Info-Gain Approach) on the aforementioned nine datasets. The analysis is presented on 20, 40, 60, 80, and 100 % of the features on 3 performance-measuring parameters Accuracy, F-Measure, and ROC.

After analyzing the performance with all kinds of datasets ranging from fewer features set to higher dimensions. It is summarized that the performance of the proposed model is improved with High-dimensional datasets. It gives good accuracy with less percentage of feature set results in less time. The experimental results revealed that the proposed model is highly recommendable to handle higher dimensional dataset not only to improve the accuracy but also to reduce the compilation time.