INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Feature Selection Techniques in Learning Algorithms to Predict Truthful Data

P Usha[1], **M P Anuradha**[2]*

1 Assistant Professor, Department of Information technology, Bishop Heber College, Affiliated to Bharathidasan University, Tiruchirappalli, 620 024, Tamil Nadu, India
2 Assistant Professor, Department of Computer Science, Bishop Heber College, Affiliated to Bharathidasan University, Tiruchirappalli, 620 024, Tamil Nadu, India

OPEN ACCESS

Received: 30-10-2022

Accepted: 08-02-2023

Published: 11-03-2023

Citation: Usha P, Anuradha MP (2023) Feature Selection Techniques in Learning Algorithms to Predict Truthful Data. Indian Journal of Science and Technology 16(10): 744-755. https://doi.org/10.17485/IJST/v16i10.2102

* **Corresponding author**.

mpanuradha.cs@bhc.edu.in

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: This review focuses on various feature selection process, strategy, and methods such as filter, wrapper and embedded algorithms and its advantages and disadvantages are presented. **Methods**: The algorithms such as Mutual Information Gain (MIG), Chi-Square (CS) and Recursive Feature Elimination (RFE) are used to select features. In this review, two benchmark datasets: Breast cancer and Diabetes are used. **Findings**: To improve the efficiency, selection of appropriate feature selection methods and algorithms are most important. To measure the performance of these selected features Random Forest model used as classifiers and compared with Support Vector Machine and Decision Tree models. Filter method and algorithm selects up to 15 features out of 17 for diabetes dataset with 89 % to 98 % of accuracy. For breast cancer dataset, up to 28 features out of 31 features selected with 98.5 % of accuracy. Wrapper method RFE selects 14 features from 17 for diabetes and 10 out of 31 features selected for breast cancer. This RFE method shows up to 98.25 % of accuracy for diabetes and 99.20% of accuracy for breast cancer. **Novelty:** Feature selection techniques help to improve the performance, efficiency and decrease the storage and processing time and build a better model for further process in prediction. The proper feature selection helps to diagnose diseases at an earlier stage and improve the survival of human beings.

Keywords: Mutual Information Gain; ChiSquare; Recursive Feature Elimination; Support Vector Machine; Random Forest; Decision Tree

1 Introduction

The real-world data may contain a lot of redundant, noisy and irrelevant features. The features which do not provide any useful information are called irrelevant features. The features which exhibit the meaning can be represented in different names as redundant features. The features that are not related to dataset or class variables are called nosy data. Eliminating these unrelated and redundant features by Feature Selection methods reduces the computation time, cost and storage. It is used to avoid degradation of learning performance and substantial loss of information and helps to develop accuracy in classification. Feature selection methods are used to reduce the dimensions of a dataset and lead to avoid overfitting of data. The feature selection, also referred to as attribute selection is also used for partitioning data into the individual classes.

Information is collected in many fields like business, hospitals, telecommunication, finance and marketing etc. that can be stored in the form of datasets. All the information that is stored in electronic form should be mined so that the relationship among data is identified and useful models can be built for interaction among data. Usually, the ground level data may contain noisy, irrelevant and redundant data. For example, the medical images dataset may contain impure data due to defects in medical imaging devices. These impurities of data in the medical field reflected in further processes will lead to improper decision-making of patient's health. Thus, in the early stage itself data must be pre-processed.

Data collected may have the form of both structured and unstructured can be represented as numeric format. To get purified data for building a model can use both feature selection and dimensionality reduction methods. Feature selection methods are used to select and exclude features without changing the features whereas Dimensionality reduction transforms the features into lower dimensions and creates an entirely new feature as input. A feature subset can be classified as: 1) Irrelevant and noisy 2) Weakly relevant and Redundant 3) Non- redundant and weakly relevant 4) Most or strongly relevant. Feature selection methods always prefer to select strongly relevant and non-redundant data from large datasets.

The aim of the feature selection method is to identify the subset of features which are meaningful from a large number of collected features. Feature selection models are very useful because: 1) It makes machine learning algorithms to train a model faster. 2) It makes the model easy to interpret and reduces the complexity of a dataset. 3) It improves the performance of a model by means of accuracy when right features are selected. 4) A proper feature selection method is used to avoid overfitting problems.

1.1 Feature selection classification

Feature selection is the process of selecting most relevant and significant features from large given datasets. A feature is a measurable property in a dataset thus it should be observable while removing in the selection process. Feature Selection techniques classified as follows and shown in Figure 1.

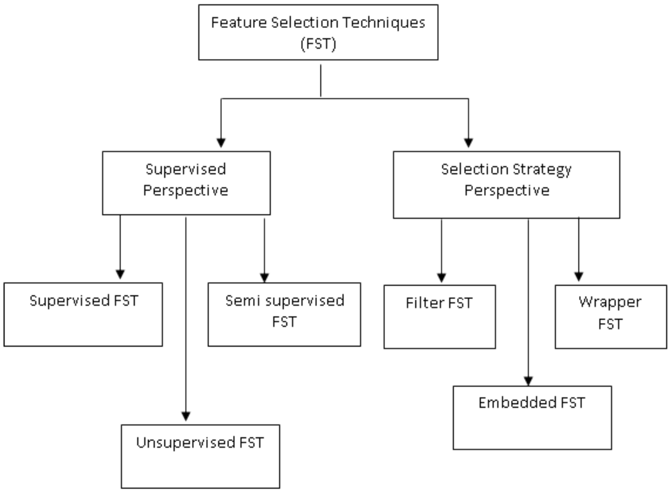

Fig 1. Feature Selection techniques classified based on supervised and selection strategy perspective based. It can be further classified and explained below

Based on supervision perspective FST can be classified as Supervised, Unsupervised and Semi supervised FST. According to the survey(1) supervised feature selection technique uses the labeled data and it is the most common and earliest section process. This technique utilizes the target variable such as search methods to remove the irrelevant variables. The process of labeling the data is costly and it is a challenging one and also leads to unreliable data. This method is sometimes forced to select irrelevant features and unintentionally remove relevant features. Supervised FST is mostly suitable for regression and classification related problems. Generally, the dataset can be split into two categories such as training and testing dataset. Training dataset highly depends on selected features.

Based on the review[2] unsupervised feature selection technique removes the redundant values using correlation methods and ignores the target variables. This technique does not require any prior knowledge on labeled data. It is an unbiased method since there is no need to categorize samples. Unsupervised FST provides an efficient way to extract the unknown features for classification of different disease types. The main issue in unsupervised FST is it neglects the useful correlation among different features and it sometimes uses mathematical concepts which are not universally accepted for all data. This method is more suitable for clustering problems.

The semi supervised FST can work with both supervised and unsupervised FST that means it can work with both labeled and unlabeled data. In this technique to maximize the margin between features labeled data can be used. To find the geometrical structure of feature unlabeled data can be used.

Based on interaction with classification or regression or clustering models, FST can be categorized as Filter, Wrapper and Embedded methods. A filter[3] FST is a supervised method which uses statistical techniques to evaluate the relationship between target and input variables. Filter method uses ranking technique to select features to build a classification model in pre-processing steps to filter a less relevant information. Ranking techniques are used to select variables and the selection process is independent of the classifier. Filter process can be done in two steps. They are 1) Rank feature 2) Filtered out low ranked features. Based on the performance values of a classifier model, the best subset can be chosen. Figure 2 shows the process of the filter method.

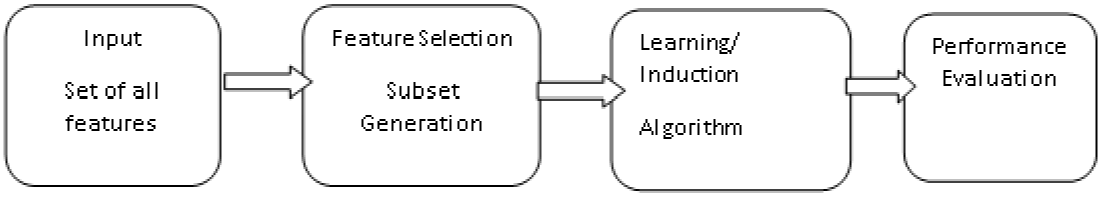

Fig 2. Process of Filter method- The dataset that contains all features can be given as input to filter based algorithms (Information gain, Gain ratio, Gini index etc.,). The extracted features performance can be analyzed with a learning algorithm (Support vector machine, K nearest neighbor, Naïve bayes and Decision tree etc.)

Wrapper feature selection method creates a model with a different subset of all input features. According to the performance metric the best model is chosen. This method requires a prior learning algorithm. Wrapper feature selection process has two steps. 1) Make a search to generate a subset. 2) Evaluate the subset by learning algorithms. Figure 3 shows the process of the wrapper method.

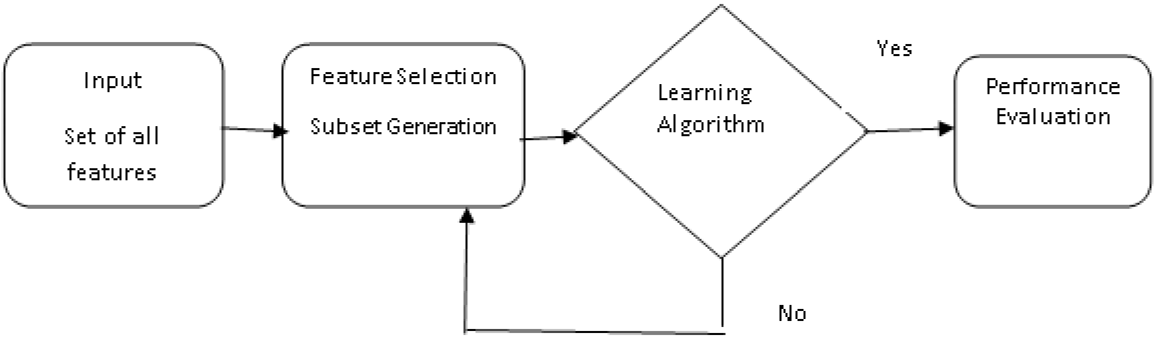

Fig 3. Process of Wrapper method- The dataset that contains all features can be given as input to wrapper based algorithms (Forward selection, Backward selection and recursive feature elimination etc.,). If the performance of extracted features with the learning algorithm (Support vector machine, K nearest neighbor, Naïve bayes and Decision tree etc.,) is satisfied performance can be evaluated otherwise the FS will continue until to get exact features

An embedded feature selection is an optimal feature generation method. In this method feature selection and learning are performed simultaneously. In the training phase itself the features are selected. This method provides the features with the highest accuracy of the model. Figure 4 shows the process of the embedded method. Every feature selection technique has pros and cons.

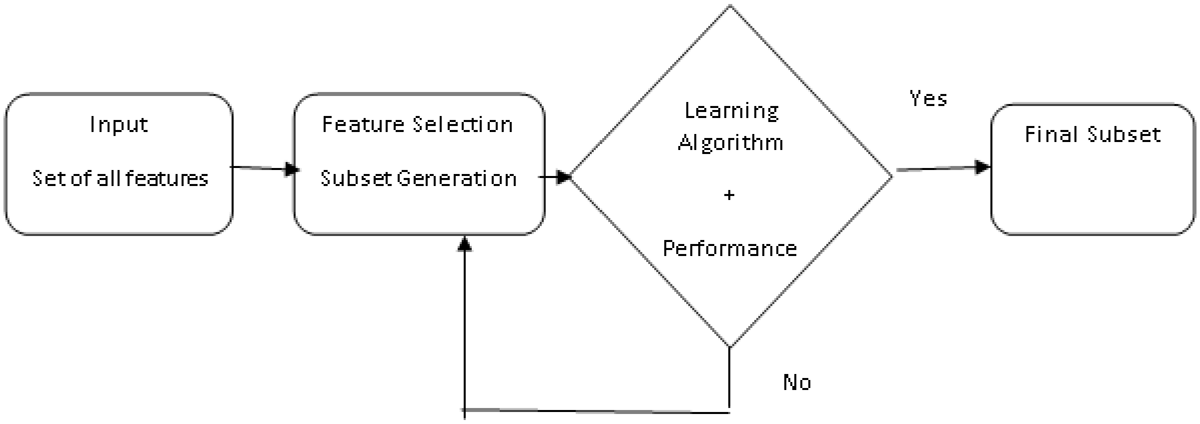

Fig 4. Process of Embedded method- It combines filter and wrapper methods of FS

1.2 Feature selection algorithm

A feature selection algorithm[4] proposes a new feature subset based on the combination of search techniques and evaluation measure depending on the feature. This algorithm is used to minimize the error rate. The evaluation metrics are important to analyze the performance of the classifier model. Thus based on evaluation metrics, feature selection algorithms can be categorized as Filter, wrapper and embedded methods. Table 1 shows the different feature selection algorithm and its pros and cons.

Table 1. Pros and Cons of Feature Selection Algorithm

Method	Algorithm	Purpose	Advantages	Disadvantages
Filter FS algorithm	$\chi 2$ test	Used to reject null hypothesis.	Test association of variables. Test undependability of data. Determine the difference between expected and observed feature values.	Use numerical data. Scale of data influences the results. Percentage cannot be used.
	Euclidean Distance	To calculate correlation between features.	Easy calculation. Time complexity is low.	Not suited for statistical problems. Depends on geometric (distance) value.
	T test	To the test null hypothesis by means.	Robustness. Easy to interpret.	Need multiple comparisons.
	Information Gain	To test the relevance of features.	Redundancy eliminated	Take la arge number of distinct features.
	Correlation based Feature Selection (CFS)	To rank features or attributes.	Eliminate more irrelevant features. Better accuracy. Less complexity.	Expensive in computation. Slower and less scalable.
	Markov Blanket Filter (MBF)	To remove irrelevant features.	Simple and fast. Independent of classification algorithm.	Feature ignorance leads to poor performance in classification.
	Fast Correlation based Feature Selection (FCBF)	To reduce redundancy in selected features.	Select higher precedence feature. Efficient.	Less scalable. Keep track of dominant features.
Wrapper	Genetic Algorithm	To detect more useful features form a large dataset included in AI.	Perform better than traditional methods. Manage dataset with many features. Easily implemented.	High computational cost. Takes long time.
	Recursive Feature Elimination (RFE)	To find the best or worst feature based on rank or score of performance metrics.	Build model with optimal solution.	Infeasible for larger dataset. Cross validation leads to high computational cost.

Continued on next page

Table 1 continued

	Sequential Selection Algorithm	To select features for the classifiers.	Interact with the classifier model. Dependency among features can be considered.	Overfitting problem occurred. Classifier dependent.
Embedded	LASSO	To enhance the interoperability and prediction accuracy.	Support large datasets. Relevance of feature is considered. Simple model and small set of key predictors. Consider all features.	Classifier dependent. Penalty function changes affect the performance. Identification of small features is problematic.
	Decision Tree	To identify the best feature for classification.	Feature scaling is not necessary. Interact with features.	Output depends on the classifier model. Suited for categorical output. Not suitable for small datasets.

1.3 Feature selection procedure

Feature selection is the process to find relevant information from a large dataset in order to obtain the best performance metrics. The following steps are involved in the feature selection process[5] and shown in Figure 5.

1. Search direction
2. Determine search strategy
3. Evaluation criteria
4. Stopping criteria
5. Validate results

1. Search direction is the first phase in feature selection which is used to find the starting point. The searching process can be done in three ways. 1) In forward searching, the search can start with an empty set and add the new features recursively in every iteration. 2) In backward searching, search can start with a full set of features and in every iteration features are removed until they reach the empty set. 3) In random searching, every iteration feature can be added as well as removed, which means it combines both forward and backward selection processes. After finding the search direction the next step, to determine the search strategy is carried out.

2. Feature selection strategies are classified as follows. 1) Forward Sequential Selection (FSS) is used to ignore the insignificant and irrelevant features and obtain an optimal subset. 2) Backward Sequential Selection (BSS) includes all the features and removes the irrelevant or redundant features one by one until to get the best feature subset. Compared to FSS this method gives a better computational effect on performance. 3) Hill Climbing (HC) which combines both FSS and BSS. In this method, the stopping criteria is set earlier by defining the number of iterations to select the optimal set. The last iteration returns the best subset for the classifier model.

3. Evaluation criteria is used to evaluate the best subset which is used to determine the relevancy towards the classification model.

4. Stopping criteria is used to specify where to stop the feature selection process in order to obtain an optima subset of features. The most common stopping criteria are number of predefined features, number of predefined iterations and evaluation function.

5. Validate result is used to check whether a selected features are valid or not or otherwise to check whether the selected features are meaningful for further process. Cross validation is a widely used validation method. Validate results give the measure such as error rate, sensitivity, specificity, ROC curve, precision, F- score, Dunn index, Jaccard index etc.

A new model CHFS-BOGA[6] (Composite Hybrid Feature Selection Learning-Based Optimization of Genetic Algorithm) proposed to predict breast cancer. This new model combines the advantages of feature selection approaches such as filter (Information Gain, Gain Ratio), wrapper (Genetic algorithm) and Embedded (C4.5) with Optimized Genetic Algorithm, Principal Component Analysis and Support Vector Machine. The new model shows the highest accuracy of 98.25% when it is combined with SVM.

An enhanced method which uses an Artificial Neural Network with a two-step feature selection method proposed in[7]. In the first step, the Best First algorithm is used to select neurons in ANN. Each node is allocated with a score through an evaluation function in ANN. The BF algorithm first scans all the nodes with the best score and maintains two lists, one list for OPEN (yet to be explored is called a priority queue) and another list for CLOSED (already visited). In this paper Taguchi method provides

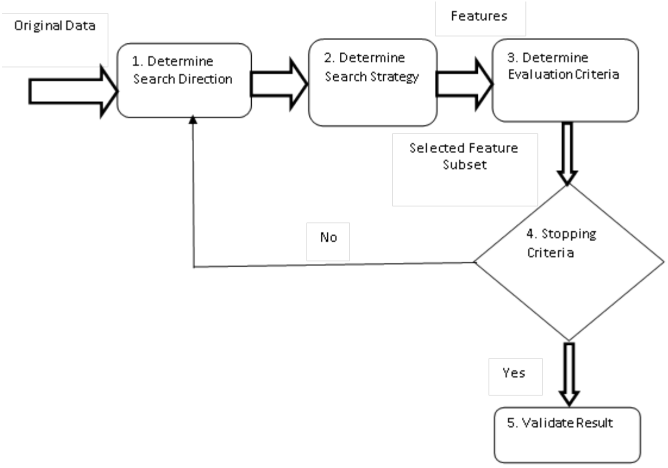
Original Data
1. Determine Search Direction
2. Determine Search Strategy
Features
3. Determine Evaluation Criteria
Selected Feature Subset
No
4. Stopping Criteria
Yes
5. Validate Result

Fig 5. FS Procedure- It shows the flow FS process. It has five steps and explained below in detail.

Orthogonal Array and Signal – to- Noise- Ratio for development and analysis. These two step feature selection methods lead to selecting the most relevant feature to build a classification model.

In[8], a feature selection method used to predict heart diseases. This paper focuses on both classification and clustering methods. Before predicting heart diseases, feature selection methods are used to extract the most relevant features for classification and grouping data. According to this paper Naïve Bayes algorithm gives better results compared to other classification models such as Discrete Transform, K- Nearest Neighbour, Genetic, Fuzzy and Neural etc.

A new hybrid algorithm which combines Simulated Annealing- Genetic Algorithm (SA-GA)[9] to find optimal feature subset for brain MR image classification. This hybrid method guarantees high computational efficiency and optimal features depend on support vector machines, greedy search, simulated annealing and genetic algorithms. SA- GA selects features without applying filter methods. This paper shows, SA used for global search ability, GA to overcome convergence issues and greedy for local search ability of feature subset generation.

An automated multimodal classification of brain tumour types using deep learning proposed in[10]. This system has five steps. 1. Linear stretching by histogram and Discrete Wavelet Transform. 2. Extract feature using CNN. 3. Select best features based on correntropy via mutual learning. 4. Find covariant features by Partial Least Square method. 5, Apply Extreme Learning machine algorithm for classification. The feature selection process is not only used to improve the classification accuracy but also reduce the computation time. This multimodal classification method shows stability in accuracy.

A new model called novel CNN in[11] uses a hyper column masking technique for brain tumor classification. The novel CNN model combines Alex Net, hyper column technique, VGG-16 RFE and SVM. To achieve deep feature Alex Net and VGG-16 is used. For enhancement Recursive Feature Elimination method is used. This proposed model is very useful in clinics for effective decision making and to reduce misdiagnosis rate. According to this model 96.77% of accuracy was obtained.

A new modern hybrid approach in[12] which includes Gray Wolf Optimizer (GWO which shows best performance in global search) and Support Vector Machine (SVM) to diagnose the tumour type whether it belongs to benign or malignant. GWO used to select parameters and SVM used to avoid overfitting. This hybrid model GWO- SVM will give the maximum accuracy of 97%.

A new hybrid data mining method[13] AP-AMBFA (Affinity Propagation- Adaptive Modified Binary Firefly Algorithm) method for diagnosing breast cancer in two phases. In the first phase pre- processing can be done through the AP method which is used to reduce the noise data. In the second phase AMBFA is used as a feature selection method and SVM used for classification. This proposed hybrid model produces 98.60 % accuracy in diagnosing breast cancer.

A 15- neuron ANN model[14] to analyze the classification accuracy in ovarian cancer detection. In this paper ANN model is used with Taguchi method which is used to select appropriate features in a dataset. Orthogonal Array is a two dimensional array which is used to find optimal neuron parameters. This model produces 98.7% of accuracy in classification and can be used as a decision support tool in the medical field. Extreme Gradient Boost method[15] for single heartbeat classification. Then a hierarchical classification method called Extreme Gradient Boosting (XGBoost) based on weight for an unbalanced heartbeat dataset which has multiple classes. In preprocessing, Recursive Feature Elimination method is used to extract features. After extraction, the XGBoost method is used in the classification stage.

In[16], three kinds of algorithms are proposed such as Weighted Gene Genetic Algorithm (WGGA), High WGGA and Low WGGAto select features from large datasets. According to this algorithm the features are classified as i) Weak or redundant features ii) Strong Relative Features iii) Unstable Features. LWGGA used to exclude features that are weak and have low parameter values. HGGA used to select the features with high parameter value (weight), which are best suited for building classification models. WGGA uses both LGGA and HGGA to select the features and reduce the time to search unimportant features in a large dataset.

A dominance based filtering approach[17] which is simple and produces fewer features for classification tasks. Mean and standard deviation for all features can be calculated and then dominance level filtering approach proceeds. Dominance level means to arrange the features according to the rank in descending order. Each individual feature has different meanings for both benign and malignant classes. In this paper different supervised algorithms such as NB, SVN and Back Propagation ANN are used to compare the performance. ANN shows higher accuracy than other classifiers.

An automated system[18] for tumor extraction and classification. This automated system consists of five steps. 1. Tumor contrast- to improve the contrast of tumour gamma contrast stretching approach is used. 2. Segmentation- done by marker-based watershed algorithm 3. Multimodal tumour extraction - to extract shape, texture and point features.4. Feature selection – chi square method. 5. Classification – Support Vector Machine. The proposed novel CNN model shows greater precision and accuracy than traditional methods.

An individual method cannot be predictable and does not guarantee that always give relevant features. It may vary depending on the application and dataset that are used in the analyzing process. Apart from individual methods hybrid or ensemble methods can be used to predict truthful data. In the existing methods, filter or wrapper or embedded feature selection with or without preprocessing mechanisms are used. In this paper two filter methods and a wrapper method is used with preprocessing to select relevant features. The filter methods give greater performance metrics and wrapper methods select relevant features with minimum time.

Key Contributions of this study listed as follows:

- Apply preprocessing technique (replace missing value with mean, standardize data) to obtain cleaned and standardized data.
- This paper demonstrates MIG, CS and RFE feature selection methods with different classifier models for Diabetes and Breast cancer datasets.
- This paper suggests the best feature selection and classifier model for the datasets used.

This paper is organized as follows: Section I provide with introduction and background study of Feature selection methods. The proposed methodology design discussed in Section II. The results obtained from the proposed design explained in Section III. Section IV concludes with results, challenges and future direction of feature selection methods.

2 Methodology

2.1 Proposed design methodologies

In the biomedical field, machine learning and deep learning algorithms play vital roles. There are many techniques and methods available in the prediction of disease types. These algorithms are most helpful for decision-making to increase the survival rate of a patient. Figure 6 shows the overview of the proposed model. In the decision-making process relevant features are most significant to predict effectiveness measures. From the study of feature selection methods, filter based methods are most suitable to select relevant features.

2.2 Pre-Processing

Dataset is a collection of records which contains attributes and values. Dataset can be collected from various sources and may have missing values, noisy data or irrelevant data. Thus pre-processing techniques are needed to clean the data. Pre-processing is a technique which can perform an earlier stage of predictive analysis. There are various kinds of pre-processing techniques available such as sampling, imputation, transform, denoising, scaling and normalization etc. In this work data is pre-processed by imputation and scaling techniques in which uniform scale is distributed for all attributes and standardized and independent features in a fixed range. In imputation, removing the missing values will be replaced by mean. The standardization or scaling value calculated using equation (1).

$$F\_New = \frac{F - Mean}{\sigma} \tag{1}$$

where F_New → new feature value, F → Feature value and σ → standard deviation.
and

$$\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N}} \tag{2}$$

where
where x_i → feature value, $\bar{x}$ → mean values of feature and N → total number of features in a dataset.

2.3 Feature selection technique

After preprocessing the dataset, Feature selection is performed in which relevant features are selected. In this paper, MIG, CS and RFE were used. The MIG feature is selected based on the gain value of an attribute. Information Gain of an attribute calculated using equation (3).

$$Information\ Gain = 1\ - E\left(S\right) \tag{3}$$

where E(S) is an entropy of an attribute and can be calculated using equation (4).

$$E\left(S\right)\ =\ \sum_{i=1}^{n} -p_i log_2 p_i \tag{4}$$

The association between categorical attributes are measured using the CS method. The association can be calculated using equation (5).

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{5}$$

where
O_i → original value and E_i → Expected value.

RFE is a wrapper based FS method which selects features based on feature score. Attributes which have the highest score are retained as selected features. The weight of an attribute is calculated using equation (6).

$$Weight\ w = \sum_{i=1}^{n} C\ f_i\ f_j \tag{6}$$

2.4 Classification models

The pre-processed and selected features are given as input to the classifier models[(19)] such Support Vector Machine (SVM), Decision tree (DT) and Random Forest (RF) models. The models performance are evaluated using different performance parameters

2.4.1 Support vector machine

Support vector machines use a hyperplane to separate data into categories. Hyperplanes used to maximize the margin between two different classes. The equation for setting plane is,

$$Y\ = a * x\ + b \tag{7}$$

The feature point that has greater or on hyper plane is named as True and if it is below it means False. The hypothesis (H) used to split classes using SVM is represented in equation (8).

$$H\left(F_i\right) = (True\ if\ ax + b\ \geq 0,\ False\ if\ ax + b\ \leq 0\} \tag{8}$$

2.4.2 Decision tree

Decision tree is a visual representation of the decision making process which uses if-then rules to predict the decision. The node in the decision tree can be split until it gets to the terminal node. In the decision tree all the true values come to the right side of the tree and false will be on the left side of the root node. The node can be splitted using Information gain or Gini index. Gini index used to split the decision nodes into various branches. It is calculated by,

$$Gini\ = 1 - \sum_{i=1}^{n} \left(p_i\right)^2 \tag{9}$$

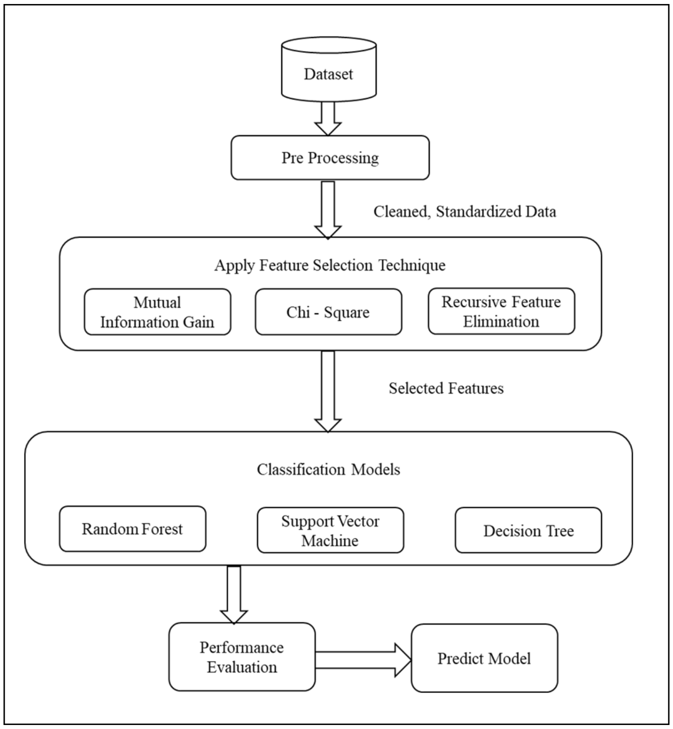

Fig 6. System Model- Dataset is preprocessed and FS techniques are applied. The extracted features performances are analyzed using classifier models.

2.4.3 Random Forest

Random Forest (RF) classifier is an ensemble technique which pools the prediction value of individual decision trees. Feature importance is calculated and features are normalized using equations (10).

$$F_{i_i} = \sum_{j:Nodes\ Split\ on\ Feature\ i} S_j\, C_j \tag{10}$$

Where $f_i$feature importance, S_j number of samples and $C_j$impurity value.

2.5 Performance Evaluation

The performance of a classifier model and feature selection technique predicted based on accuracy. Accuracy calculated using equation (11).

$$Accuracy = \frac{TP + TN}{TP + TN + Fp + FN} \tag{11}$$

where

TP → True Positive, TN → True Negative, FP → False Positive and FN → False Negative.
TP → represents predicted YES with actual YES
TN → represents predicted NO with actual NO
FP → represents predicted YES with actual NO
FN → represents predicted NO with actual YES

3 Results and Discussion

Diabetes and Breast cancer dataset downloaded from UCI repository. In previous research, pre-processing done by replacing mean or median and embedded methods are used to select features. In this work, pre-processing is done with imputation and scaling techniques which helps to get more cleaned and standardized dataset. Features are selected with filter and wrapper methods and compared with different classifier models. Table 2 shows the selected features through IG, CS and RFE.

Table 2. Selected Features

Dataset	Total Number of Features	Number of Selected Features		
		IG	CS	RFE
Diabetes	17	15	14	**14**
Breast Cancer	31	28	25	**10**

Figure 7 shows the graphical representation of selected features through filter and wrapper methods. RFE wrapper method select most relevant and optimal feature for decision-making system.

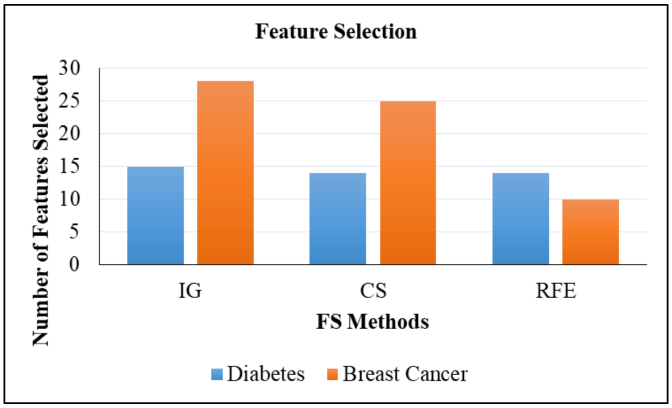

Fig 7. Features obtained through FS methods

The previous work[20] shows 93.54% of accuracy with RF classifier,[21] shows 97.36 % 0f accuracy with Nearest Neighbour (KNN) model for breast cancer detection and[22] combines MIG and sequential forward selection wrapper method to select features and shows almost 100 % of accuracy with SVM, DT, KNN and RF classifier . In[23] mean imputation technique with RF model and deep learning algorithms for predicting breast cancer gives 98.75% of accuracy. In our work the accuracy of standardized dataset through FS methods is evaluated using SVM, DT and RF classifier models. The FS results obtained and executed in Jupyter Notebook using Python and shown in Table 3 .

Table 3. Performance: Accuracy

Dataset/ FS Methods/ Classifier Model	Accuracy in %					
	Diabetes			Breast Cancer		
	SVM	DT	**RF**	SVM	DT	**RF**
IG	89	96	98	98.5	99	99
CS	89	95	98	98.2	98.5	99
RFE	98	96	**98.25**	99	99	**99.2**

The performance of FS methods are evaluated through classifier models and shown in Figure 8 . Random Forest model with RFE FS method gives better performance than other model and FS methods.

4 Conclusion

This study has discussed the feature selection procedure, different types of feature selection techniques and varieties of selection algorithms according to techniques with their merits and demerits. Considering a dataset as whole is difficult to process so it can be extracted with necessary information in order to reduce time resources etc. From this study we found, ensemble or hybrid feature selection gives better results than individual methods. Although the hybrid methods always show efficiency depending on the dataset, application area feature selection algorithms can vary. Every algorithm has its own characteristics to solve the problem. Feature selection methods and algorithms are used to enhance the performance of classifier models. Filter based methods and recursive feature elimination algorithms are very useful to select optimal subset of features in the biomedical field because of their robustness. This paper concludes, RFE feature selection methods selects optimal features such that 14 features

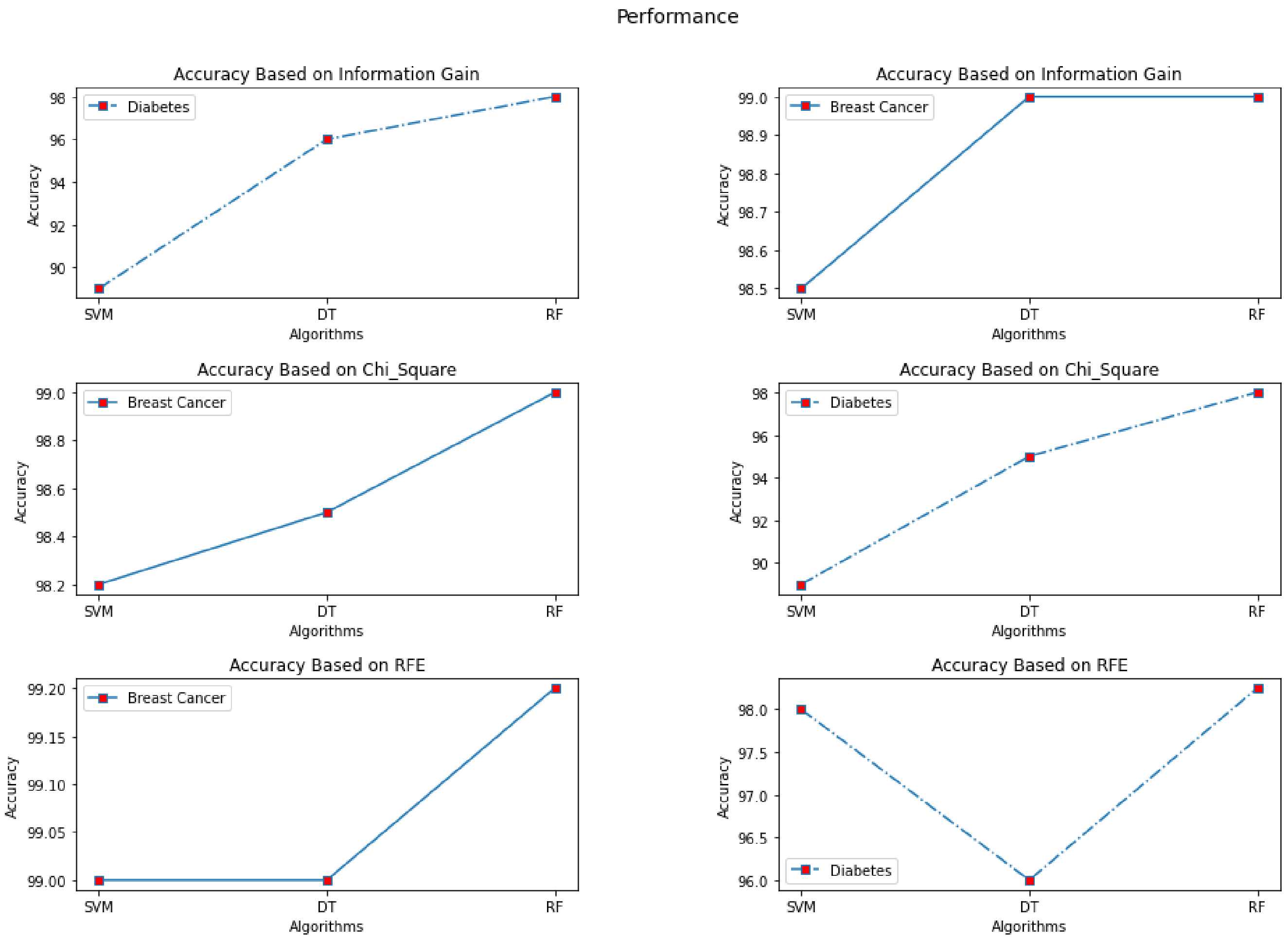

Fig 8. FS method (IG, CS and RFE) performance based on Accuracy through classifier models (SVM, DT and RF).

from 17 for diabetes and 10 out of 31 features selected for breast cancer and gives 99.2% of accuracy with RF classifier model for breast cancer and 98.25% for diabetes. We arrive at conclusion that the single method does not give better results but a hybrid approach can help to improve the efficiency and result. In future, more than one filter based feature selection method and wrapper approaches can be combined and applied to predict truthful features for further classification process.

References

1) Islam MR, Lima AA, Das SC, Mridha MF, Prodeep AR, Watanobe Y. A Comprehensive Survey on the Process, Methods, Evaluation, and Challenges of Feature Selection. *IEEE Access*. 2022;10:99595–99632. Available from: https://doi.org/10.1109/access.2022.3205618.
2) Pudjihartono N, Fadason T, Kempa-Liehr AW, O&apos;sullivan JM. A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. *Frontiers in Bioinformatics*. 2022;2:927312. Available from: https://doi.org/10.3389/fbinf.2022.927312.
3) Wang Y, Gao X, Ru X, Sun P, Wang J. A hybrid feature selection algorithm and its application in bioinformatics. *PeerJ Computer Science*. 2022;8:e933. Available from: https://doi.org/10.7717/peerj-cs.933.
4) Hessen SH, Abdul-Kader HM, Khedr AE, Salem RK. Developing Multiagent E-Learning System-Based Machine Learning and Feature Selection Techniques. *Computational Intelligence and Neuroscience*. 2022;2022:1–8. Available from: https://doi.org/10.1155/2022/2941840.
5) Venkatesh B, Anuradha J. A review of Feature Selection and its methods. 2019. Available from: https://doi.org/10.2478/cait-2019-0001.
6) Farid A, Selim A, Khater G, H. A composite hybrid feature selection learning-based optimization of Genetic Algorithm for breast cancer detection. . Available from: https://doi.org/10.20944/preprints202003.0298.v1.
7) Rahman MA, Muniyandi RC. An Enhancement in Cancer Classification Accuracy Using a Two-Step Feature Selection Method Based on Artificial Neural Networks with 15 Neurons. *Symmetry*. 2020;12(2):271. Available from: https://doi.org/10.3390/sym12020271.
8) A RA, R K. An analysis on feature selection methods, clustering and classification used in heart disease prediction –a machine learning approach. *Journal of critical reviews*. 2020;7(06). Available from: https://doi.org/10.31838/jcr.07.06.27.

9) Tang J, Wang Y, Luo Y, Fu J, Zhang Y, Li Y, et al. Computational advances of tumor marker selection and sample classification in cancer proteomics. *Computational and Structural Biotechnology Journal*. 2020;18:2012–2025. Available from: https://doi.org/10.1016/j.csbj.2020.07.009.

10) Khan MA, Ashraf I, Alhaisoni M, Damaševičius R, Scherer R, Rehman A, et al. Multimodal Brain Tumor Classification Using Deep Learning and Robust Feature Selection: A Machine Learning Application for Radiologists. *Diagnostics*. 2020;10(8):565. Available from: https://doi.org/10.3390/diagnostics10080565.

11) Toğaçar M, Cömert Z, Ergen B. Classification of brain MRI using hyper column technique with convolutional neural network and feature selection method. *Expert Systems with Applications*. 2020;149:113274–113274.

12) Badr ESA, Ahmed MS, Hagar. Optimizing Support Vector Machine using Gray Wolf Optimizer Algorithm for Breast Cancer Detection. 2019. Available from: https://www.researchgate.net/publication/337151922_Optimizing_Support_Vector_Machine_using_Gray_Wolf_Optimizer_Algorithm_for_Breast_Cancer_Detection/citations.

13) Emami N, Pakzad A. A New Knowledge-Based System for Diagnosis of Breast Cancer by a combination of the Affinity Propagation and Firefly Algorithms. *Journal of AI and Data Mining*. 2019;7(1):59–68. Available from: https://doi.org/10.22044/jadm.2018.6489.1763.

14) Rahman MA, Muniyandi RC, Islam KT, Rahman MM. Ovarian Cancer Classification Accuracy Analysis Using 15-Neuron Artificial Neural Networks Model. *2019 IEEE Student Conference on Research and Development (SCOReD)*. 2019. Available from: https://www.researchgate.net/publication/341434081_Ovarian_Cancer_Classification_Accuracy_Analysis_Using_15-Neuron_Artificial_Neural_Networks_Model.

15) Shi H, Wang H, Huang Y, Zhao L, Qin C, Liu C. A hierarchical method based on weighted extreme gradient boosting in ECG heartbeat classification. *Computer Methods and Programs in Biomedicine*. 2019;171:1–10. Available from: https://doi.org/10.1016/j.cmpb.2019.02.005.

16) Mohammed TA, Bayat O, Uçan ON, Alhayali S. Hybrid Efficient Genetic Algorithm for Big Data Feature Selection Problems. *Foundations of Science*. 2020;25(4):1009–1025. Available from: https://doi.org/10.1007/s10699-019-09588-6.

17) Atrey K, Sharma Y, Bodhey NK, Singh BK. Breast Cancer Prediction Using Dominance-based Feature Filtering Approach: A Comparative Investigation in Machine Learning Archetype. *Brazilian Archives of Biology and Technology*. 2019;62. Available from: https://doi.org/10.1590/1678-4324-2019180486.

18) Khan MA, Lali IU, Rehman A, Ishaq M, Sharif M, Saba T, et al. Brain tumor detection and classification: A framework of marker-based watershed algorithm and multilevel priority features selection. *Microscopy Research and Technique*. 2019;82(6):909–922. Available from: https://doi.org/10.1002/jemt.23238.

19) Phagwara P, India, Nidhi, Sharma B, Handa D. Building predictive model by using data mining and feature selection techniques on academic dataset. *International Journal of Modern Education and Computer Science*. 2022;14(4):16–29. Available from: https://doi.org/10.5815/ijmecs.2022.04.02.

20) Khan F, Tarimer I, Alwageed HS, Karadağ BC, Fayaz M, Abdusalomov AB, et al. Effect of Feature Selection on the Accuracy of Music Popularity Classification Using Machine Learning Algorithms. *Electronics*. 2022;11(21):3518. Available from: https://doi.org/10.3390/electronics11213518.

21) Jain S, Kumar P. Accuracy Enhancement for Breast Cancer Detection Using Classification and Feature Selection. *International Journal of Information Retrieval Research*. 2022;12(2):1–15. Available from: https://doi.org/10.4018/ijirr.299931.

22) Elemam T, Elshrkawey M. A Highly Discriminative Hybrid Feature Selection Algorithm for Cancer Diagnosis. *The Scientific World Journal*. 2022;2022:1–15. Available from: https://doi.org/10.1155/2022/1056490.

23) Gomes R, Paul N, He N, Huber AF, Jansen RJ. Application of Feature Selection and Deep Learning for Cancer Prediction Using DNA Methylation Markers. *Genes*. 2022;13(9):1557. Available from: https://doi.org/10.3390/genes13091557.