International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10s
DOI: https://doi.org/10.17762/ijritcc.v11i10s.7643
***Article Received:** 01 June 2023 **Revised:** 30 July 2023 **Accepted:** 12 August 2023*

A Comparative Analysis for Filter-Based Feature Selection Techniques with Tree-based Classification

Mahesh Parmar[1], Abhilash Sonker[2], Vikas Sejwar[3]
[1]Assistant Professor, Department of CSE
Madhav Institute of Technology & Science
Gwalior, Madya Predesh, India-474005
maheshparmar@mitsgwalior.in
[2]Assistant Professor, Department of IT
Madhav Institute of Technology & Science
Gwalior, Madya Predesh, India-474005
abhilashsonkerit@mitsgwalior.in
[3]Assistant Professor, Department of IT
Madhav Institute of Technology & Science
Gwalior, Madya Predesh, India-474005
vikassejwar@mitsgwalior.in

Abstract—The selection of features is crucial as an essential pre-processing method, used in the area of research as Data Mining, Text mining, and Image Processing. Raw datasets for machine learning, comprise a combination of multidimensional attributes which have a huge amount of size. They are used for making predictions. If these datasets are used for classification, due to the majority of the presence of features that are inconsistent and redundant, it occupies more resources according to time and produces incorrect results and effects on the classification. With the intention of improving the efficiency and performance of the classification, these features have to be eliminated. A variety of feature subset selection methods had been presented to find and eliminate as many redundant and useless features as feasible. A comparative analysis for filter-based feature selection techniques with tree-based classification is done in this research work. Several feature selection techniques and classifiers are applied to different datasets using the Weka Tool. In this comparative analysis, we evaluated the performance of six different feature selection techniques and their effects on decision tree classifiers using 10-fold cross-validation on three datasets. After the analysis of the result, It has been found that the feature selection method ChiSquaredAttributeEval + Ranker search with Random Forest classifier beats other methods for effective and efficient evaluation and it is applicable to numerous real datasets in several application domains

Keywords- Feature selection Methods, Search Method, Decision tree, j48, REP, Random Forest Attribute evaluator, weka tool

I. INTRODUCTION

Feature selection techniques have played an important role in numerous research areas including, statistics, pattern identification, text mining, machine vision, biomedical, and data mining communities. Data collected by researchers from various resources like websites, social media, Twitter, blogs, etc., are unstructured data. The machine learning models cannot be fed unstructured data directly. Data pre-processing technique is used in the first phase of machine learning, which consists of data reduction, integration, transformation, and cleaning. A pre-processing technique is feature selection that chooses the optimum combination of features to capture the dataset's relevant information, which allows one to construct effective machine learning models for the research phenomena. Basically, the feature selection method finds the minimum subset of features and removes irrelevant and redundant information from the dataset and significantly decreases the time needed for the modeling and increases the overall quality of the data. When the machine learning model is built It offers a number of advantages, including decreasing overfitting, increasing accuracy, and reducing the learning model's training period.

With respect to the selection and evaluation measure, the filter and wrapper method can be used for feature selection. Filter method is independent of the classification algorithm. It is used to identify the subset of features and assess the quality of these subsets[1] [23]. Predefined classification is used in the wrapper method, this classifier evaluates the subsets of selected features. Single and subset evaluation categories are used to categorize feature selection techniques depending on how the features are evaluated. When Single evaluation is used [2], it is referred to as feature ranking or feature weighting, each feature

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10s
DOI: https://doi.org/10.17762/ijritcc.v11i10s.7643
***Article Received:** 01 June 2023 **Revised:** 30 July 2023 **Accepted:** 12 August 2023*

is evaluated separately by giving a weight based on how important it is, however when using subset evaluation, each subset of characteristics is assessed separately [3][4].

The wrapper method gives good results as compared to the filter method because Wrapper approaches assess the "usefulness" of features from the selected subsets of features based on the performance of the classifier, but it is more expensive than filter methods [2][3]. The significance of a variable has been defined and measured in several publications [5]–[8] using a variety of different measures. For the process of feature reduction and selection, there are many models available today but only a few models are suitable for real-time application [2]. Therefore, it is important to research whether attribute selection methods for a standard database are appropriate. In this paper, we determine how to perform feature subset selection on standard datasets using Weka Tool. Basically, we applied six attribute evaluator, CorrelationAttributeEval, CfsSubsetEval, ChiSquaredAttributeEval, InfoGainAttributeEval, GainRatio AttributeEval and WrapperSubsetEval and three searching method, bestFirst, Ranker and genetic search. The performance is evaluated after applying the evaluator using 10 cross validations with different decision tree methods, J48, Random Forest and BEP for different datasets, Vehicle, Vote and Credit to predict the relevant features

II. RELATED WORK

The variable Ranking technique is used in the filter method, the main criteria for selecting variables from datasets is by order. Due to its simplicity and reported good success, it is used for practical applications. A score of the variable is calculated by a suitable ranking criterion and the threshold value is decided to remove variables that have scores below the threshold and filter out the less relevant variables. Ranking techniques are filter techniques since they are used prior to classification to filter out the irrelevant factors. Unique features have some fundamental characteristics, including the ability to distinguish between other classes by containing relevant and useful information about them. [3].

A lot of search algorithms are used to find and evaluate subsets of variables for the feature selection subsets of variables maximizing the objective function which computes the classification performance. The Branch and Bound method [5][6], finds a different subset for feature selection from a given feature selection number using a tree structure. Sequential search and evolutionary algorithm, Genetic Algorithm (GA) [7] or Particle Swarm Optimization (PSO) is a simplified algorithm that produces computationally feasible optimum results for a higher number of features. Exhaustive search methods are becoming computationally intensive for larger datasets.

The Sequential Feature Selection (SFS) algorithm, and Sequential Floating Forward Selection (SFFS) [8][9] algorithm produce value for the objective function. Due to an additional backtracking step, SFFS is more versatile than naive SFS. For embedded methods [10] [11] [12], computation time taken up by reclassification of different subsets can be reduced which is done in wrapper methods. The primary strategy is to include feature selection in the training procedure. Here Table 1 mentioned the feature selection method which has been used in various kinds of researches.

TABLE 1 REVIEW ON FEATURE SECTION METHOD

Paper Citation	Datasets	Data size	Data selection Method	Classifier	Finding
I. M. El-hasnony and et al. [13]	Breast cancer dataset	Attribute 10, Instances 699,	rough set , gain ratio, principal, and correlation feature selection	Decision-Tree	Fuzzy rough set feature selection shows better results as compared to other techniques.
Kohavi, Ron et. al. [14]	breast cancer, cleve, crx,, DNA	Instances 699, Instances 303, Instances 690, Instances 2000	Relief	Naive-Bayes and decision-tree	Compare Relief with the wrapper technique with feature selection and induction method, without feature subset selection
Kwak et. al. [15]	IBM Datasets	Attribute 09, instances 1000	Taguchi Method in Feature Selection (TMFS)	MLP	It is predicted that MIFS-U with TMFS will significantly increase performance.
Feedback, Share et. al. [16]	Customer Relationship Management	Instances 200	Relief-F	Support Vector Machines and K-NN	Compare of various feature selection methods and how they impact different domain-specific classification algorithms
Karegowda and Asha Gowda [17]	Pima Indians Diabetes Database	Attribute 8, Instances 768	Genetic search with Correlation based	featureC4.5 decision tree with gain ratio	Experimental results show that the CFS selected feature subset has enhanced the BPN and RBF classification accuracy.
Zhao, Zheng Wang et. al. [18]	RELATHE	Feature 4322, Instance 1427	Relief	SVM	Discovered that the least square formulations of many learning models, including PCA, LDA, and SVM, may be used to connect the SPFS formulation to those models.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10s
DOI: https://doi.org/10.17762/ijritcc.v11i10s.7643
***Article Received:** 01 June 2023 **Revised:** 30 July 2023 **Accepted:** 12 August 2023*

Peng, Hanchuan Long et.al.[19]	HDR MultiFeat, Arrhythmia	Feature 649, Sample 2000	Minimal redundancy maximal relevance (mRMR)	NB, SVM	Using mRMR feature selection, classification accuracy can be greatly increased.
Osanaiye, Opeyemi Cai et. Al. [20]	NSL-KDD dataset features	Feature 41, Instances 60167	Ensemble IG, gain ratio, ReliefF and chi-squared	J48 decision tree	Performance evaluation of EMFFS approach by NSL-KDD, dataset showed it outperforms is better than other separate filter based feature selection approach with the decision tree classifier.
Ileberi, Emmanuel et. al. [21]	Credit card transactions that	Feature 30, Instances 284807	GA-based	ANN	Using GA-selected attribute method, the experimental results demonstrated that the GA-RF captured an overall optimal accuracy of 99.98%..

III. FEATURE SELECTION METHODS

In Science, a large number of variables are used in the datasets, and the development and training of models can be slow as well as requires a large amount of memory. Moreover, if the variables are not relevant to the desired variable then the performance of the model can be degraded. Basically supervised and unsupervised feature selection methods are used to remove the irrelevant and redundant variable from the data, in the supervised method the features are selected based on the target variable whereas in the unsupervised feature selection method ignore the target variable and remove redundant variable using correlation [24]. Here, filter, wrapper, and embedded feature selection methods are used to remove redundant variables.

A. Filter methods

Filtering techniques capture the fundamental qualities of both features evaluated by univariate statistics. Cross-validation performance should not be the main priority. Compared to wrapper approaches, these strategies are faster and more effective in terms of computation. It is computationally more inexpensive than other feature selection approaches when dealing with high-dimensional data [3] [26].

B. Wrapper Methods:

In order to evaluate the quality of each feasible feature subset, wrappers need a method for scanning the space of all possible feature subsets and training and assessing classifiers with each feature subset. It uses a greedy search approach to compare each and every potential feature combination to the evaluation criterion. In most situations, wrapper strategies outperform filter strategies in terms of the precision of predictions.

C. Embedded Methods:

These methods take feature relations into consideration while maintaining an inexpensive computational cost, incorporating the advantages of the wrapper and filter approaches. Embedded methods are iterative include that they pay close attention to each stage of the model-training process and carefully pick out the features that are most beneficial for training for that stage.

D. Correlation Based Feature Selection:

Measuring correlation between features and classes as well as between features and other features is one of the well-known strategies for choosing the most pertinent features. In statistics, correlation is more formally known as Pearson's correlation coefficient. The inputs of the number of features k and the classes C, CFS determined the relevance of the features subset in following equation.

$$Mrits = \frac{Kr_{nc}}{\sqrt{n+(n-1)r_{nn}}} \quad \text{..... .....(1)}$$

Where, Mrits,= Relevance feature, r_{nc} = Average Linear Correlation Coefficient between the Class and the Feature, r_{nn} = Average Linear correlation coefficient between different features

E. Information Gain Based Feature Selection:

A strategy for selecting features based on information gain determine the entropy or information gain for each attribute for the output variable. Entry values range from 0 to 1, where 0 represents no information and 1 represents all available information. The attributes that contribute the most information and have the highest information gain value are chosen, while those that don't offer much information or have a low score can be eliminated. Calculate the information gain for the attributes given in the simplified discernibility function according to the definition of the information gain is

$$Gain(G_j) = E(R_j) - E(G_j)$$

$$E(R) = \sum_{i=1}^{n} R_i \, Log_2 R_i$$

$$= -\frac{R_1}{R} Log_2 \frac{R_1}{R} - \frac{R_2}{R} Log_2 \frac{R_2}{R} \ldots\ldots\ldots\ldots \frac{R_m}{R} Log_2 \frac{R_m}{R} \quad \text{...(2)}$$

Here R_i is the ratio of the dataset's conditional attribute R. While G_j have $|G_j|$ types of attribute values and

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10s
DOI: https://doi.org/10.17762/ijritcc.v11i10s.7643
***Article Received:** 01 June 2023 **Revised:** 30 July 2023 **Accepted:** 12 August 2023*

conditional attribute R_i is a partitions set of R using attribute G_j . The value of information $E(G_j)$ as defined

$$E(G_j) = \sum_{i=1}^{n} I_j * E(Y_j) \quad \text{..... (3)}$$

F. Algorithm 1: Feature Selection Procedure

According to Figure 1, the feature selection procedure incorporates the following four steps:

Step-I: Subset Generation: At this step, using a specific search technique, subsets of features are created for the evaluation

Step-II: Subset Evaluation: In this step, using evaluation function the caliber of the subset is evaluated which is produced by generation criterion. The subset produced is compared to the prior best and, if better, is replaced with it

Step-III: Stopping Criterion: In this step, a stop condition is used when achieving an ideal subset of features, halting at a present number of features or iterations

Step-IV: Result Validation: At this step, comparing the resulting subset with the previously discovered using The validity of the subset of the chosen features is tested using a number of different methods.

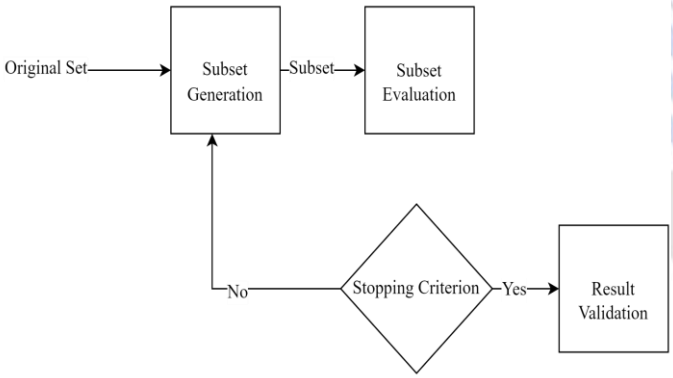

Figure 1 Selection Process [22]

G. Algorithm 2: Decision Tree

Step I: Select a suitable attribute that distinguishes the output attribute values the best.

Step II: For each value of the selected attribute, different branches of the tree is created.

Step III: To reflect the attribute values of the selected node, divide the instances into subgroups.

Step IV: If we meet the following criteria, stop selecting attributes for each subgroup.

a) When every member of a subgroup has the same value for the output attribute and the branch on the current path is labelled with the required value, the attribute selection process for the current path is stopped.
b) There is only one node in the subgroup, or no more distinguishing characteristics can be discovered. The branch should be labelled with the output value that the great majority of the remaining instances see, as in (a).

Step V: Each subgroup generated in (Step 3) that hasn't been labelled as terminal then repeat the process above.

IV. EPERIMENTAL RESULT AND DISCUSSIONS

In this research, we used the WEKA tool to implement the feature selection method. This tool is open-source software which is written in java programming. Various machine learning algorithms are included in this tool such as data preparation, classification, regression, clustering, association rules, and visualization. For evaluating the subset of selected features, learning algorithm J48 decision trees are used. Basically WEKA tool contains two components, Feature evaluator and Search method. feature subsets are evaluated using the feature evaluator and the space of feature is searched by search method [27].

A. Feature evaluator

1. ***CorrelationAttributeEval:*** It provides us with the attributes rankings from highest to lowest and displays the rank number as well. The attribute selection approach is effective since it produces results without the support of any other algorithm, including J48 and others.
2. ***InfoGainAttributeEval :*** *To determine the value of a specific attribute, one measures the information obtained in relation to the class. InfoGain(Class,Attribute) = H(Class) - H(Class Attribute).*
3. ***WrapperSubsetEval:*** *Utilises a learning technique to assess attribute sets. The learning strategy's accuracy for a specific set of features is determined through cross-validation.*
4. ***CfsSubsetEval:*** *Considers the individual applicability of each feature as well as the degree of overlap between them to determine the importance of a subset of attributes. It is preferable to use subsets of attributes that have minimal inter correlation with other attributes and high correlation with the class.*
5. ***ChiSquaredAttributeEval:*** *Determines the value of a specified attribute by computing the chi-squared statistic's value with respect to the class.*
6. ***GainRatioAttributeEval:*** *Evaluates an attribute's value by calculating the gain ratio relevant to the class [28].*

GainR (Class, Attribute) = (H (Class) _ H(Class | Attribute)) / H (Attribute)

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10s
DOI: https://doi.org/10.17762/ijritcc.v11i10s.7643
***Article Received:** 01 June 2023 **Revised:** 30 July 2023 **Accepted:** 12 August 2023*

B. Search Methods:

The best possible set of features is found using search methods by searching the entire set of all potential features. In this work, four search techniques BestFirst, GeneticSearch, GreedyStepwise, and Ranker—that are used which is available in Weka are used for comparison purposes.

1. ***BestFirst:*** To search the space of attribute subsets, best fit search uses greedy hill climbing with a backtracking facility. Best first can begin at any point and search in both ways, or it can begin with a set of attributes that is empty and search ahead, or it can begin with a set of attributes that is complete and search backward.
2. ***GeneticSearch:*** It uses a simple genetic algorithm to conduct the search [29].
3. ***GreedyStepwise:*** It performs a greedy search in the space of attribute subsets, moving either forward or backward. Start at any place in the space or with all attributes. Stops when the evaluation decreases as a result of adding or removing any remaining attributes. It is also possible to construct a ranked list of attributes.
4. ***Ranker:*** According to each attribute's separate ratings, it provides ranks to all attributes. It is used combination of attribute evaluators (Chisquare, GainRatio, InfoGainetc).

C. Datasets

Four datasets are evaluated in this research. A vehicle dataset, with a set of attributes that were extracted from the silhouette, assigns the particular silhouette to one of four different vehicle categories. Vote dataset: Describe the actual votes according to each member of the US House of Representatives and Recognize them as Republicans or Democrats. And Credit dataset, based on a variety of attributes, this dataset classifies individuals to the categories of good or bad credit risks. The details regarding the datasets is shown in table 2.

TABLE 2 DATASETS DESCRIPTION

S.No	Datasets	Features	Instances	Classes	Dataset Sources
1	Vehicle	18	946	4	[30]
2	Vote	16	435	2	[31]
3	Credit	20	1000	2	[32]

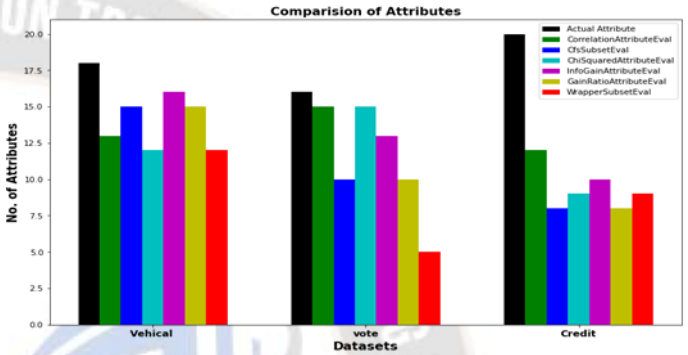

Figure 2 Comparison of Attribute

Figure 2 shows comparison among the vehical, Vote and Credit datasets attributes. Here we seen, 12 sequence of attributes are generated by wrapperSubsetEval feature selection method on the veical datasets, 5 sequence of attributes are generated by wrapperSubsetEval feature selection method on the vote datasets, and 8 sequence of attributes are generated by GainRatioAttributeEval and CfsSubsetEval feature selection methods on the Credit datasets.

TABLE 3 COMPARISON DIFFERENT FEATURE SECTION METHODS WITH CLASSIFICATION ACCURACY

Datasets	Attribute evaluator	Search method	Actual Attribute	Selected Attribute	Selected Attribute sequence	Classification Accuracy			
						Accuracy before applying the evaluator	Accuracy after applying the evaluator using 10 fold cross validation		
							J48 classifier	Random Forest Classifier	REP Tree classifier
Vehicle	CorrelationAttributeEval	Best First	18	13	3,8,7,12,11,9,4,14,1,18,16,13,2	72.4586	**73.1678**	**73.1678**	**68.4397**
	CfsSubsetEval	Best First	18	15	7,8,11,9,3,6,2,1,4,13,10,14,17,18,5	72.4586	73.234	74.532	74.8923
	ChiSquaredAttributeEval	Ranker	18	12	2,17,12,14, 7,8,11,9,3,6,2,1	72.4586	73.235	74.457	72.231
	InfoGainAttributeEval	Ranker	18	16	12,7,8,11,9,3,6,2,1,4,13,10,14,17,18,5	72.4586	74.4681	**76.3593**	73.2861
	GainRatioAttributeEval:	Ranker	18	15	11,17,14,10,13,4,1,2,6,3,9,8,7,5,12	72.4586	72.783	70.342	73.432
	WrapperSubsetEval	Genetic Search	18	12	3,4,5,6,7,8,10,12,13,15,17,18	72.4586	72.8132	70.5674	73.8771
	CorrelationAttributeEval	Best First	16	15	4,3,5,12,9,8,14,13,15,7,6,2,16,10,11	96.3218	96.3218	95.8621	95.4023

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10s
DOI: https://doi.org/10.17762/ijritcc.v11i10s.7643
***Article Received:** 01 June 2023 **Revised:** 30 July 2023 **Accepted:** 12 August 2023*

	CfsSubsetEval	Best First	16	10	3,5,12,14,8,9,7,6,1,18	96.3218	97.453	96.564	95.456
	ChiSquaredAttributeEval	Ranker	16	15	12,14,8,9,13,15,7,6,1,11, 4,3,5,10,2	96.3218	95.876	**98.345**	97.312
	InfoGainAttributeEval	Ranker	16	13	4,3,5,12,14,8,9,13,15,7,6,1,11	96.3218	96.3218	95.8621	95.4023
	GainRatioAttributeEval:	Ranker	16	10	5,12,9,8,14,13,15,7,6,2	96.3218	95.456	96.765	95.678
Vote	WrapperSubsetEval	Genetic Search	16	5	3,4,7,9,11	96.3218	96.3218	96.7816	95.6322
	CorrelationAttributeEval	Best First	20	12	1,2,5,6,15,14,13,3,20,4,8,9	70.5	72.5	75.6	71.7
	CfsSubsetEval	Best First	20	8	2,5,6,15,14,13,3,20	70.5	69.6	71.8	73.6
	ChiSquaredAttributeEval	Ranker	20	9	1,3,2,6,4,5,12,7,	70.5	74.8	**76.8**	73.5
	InfoGainAttributeEval	Ranker	20	10	1,3,2,6,4,5,12,7,15,13	70.5	72.4	76.3	72.6
Credit	GainRatioAttributeEval:	Ranker	20	8	2,6,4,5,12,7,15,13	70.5	72.3	74.5	73.8
	WrapperSubsetEval	Genetic Search	20	9	1,2,3,5,6,10,13,16,20	70.5	73.1	73.8	73.6

In this paper, we used the machine learning tool WEKA for comparative analysis of different feature selection algorithms, filter, wrapper and embedded. We applied six attribute evaluator, CorrelationAttributeEval, CfsSubsetEval, ChiSquaredAttributeEval,InfoGainAttributeEval,GainRatio AttributeEval and WrapperSubsetEval and three searching method, bestFirst, Ranker and genetic search. The performance is evaluated after applying the evaluator using 10 cross validations with different decision tree methods, J48, Random Forest and BEP for different datasets, Vehicle, Vote and Credit.

According to Table 3, for vehicle dataset, we observed that InfoGainAttributeEval + Ranker attribute evaluator with J48 classifier has greater accuracy (74.457) as compared to others. For the Vote dataset, we observed that CfsSubsetEval + Best First attribute evaluator with J48 classifier has greater accuracy (97.453) as compared to others. Similarly, For Credit data set , ChiSquaredAttributeEval + Ranker attribute evaluator with J48 classifier have greater accuracy (74.8) as compared to others. Now we applied each attribute evaluator using 10 cross validations with Random Forest Classifier. We observed that InfoGainAttributeEval + Ranker attribute evaluator has greater accuracy (76.3593) as compared to others. For the Vote dataset, we observed that ChiSquaredAttributeEval + Ranker attribute evaluator has greater accuracy (98.345) as compared to others. Similarly, For Credit data set , ChiSquaredAttributeEval + Ranker attribute evaluator has greater accuracy (76.8) as compared to others.

Similarly, when we applied each attribute evaluator using 10 cross validations with REP tree Classifier. We observed that WrapperSubsetEval + Genetic Search attribute evaluator have greater accuracy (73.8771) as compared to others. For the Vote dataset, we observed that ChiSquaredAttributeEval + Ranker attribute evaluator has greater accuracy (97.312) as compared to others. Similarly, For Credit data set , both WrapperSubsetEval + Genetic Search and CfsSubsetEval + Best attribute evaluator have greater accuracy (73.6) as compared to others.

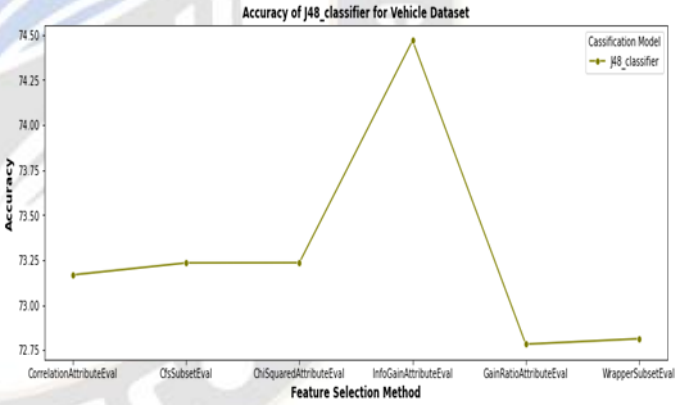

Figure 3 Accuracy of J48 Classifier for Vehicle Dataset

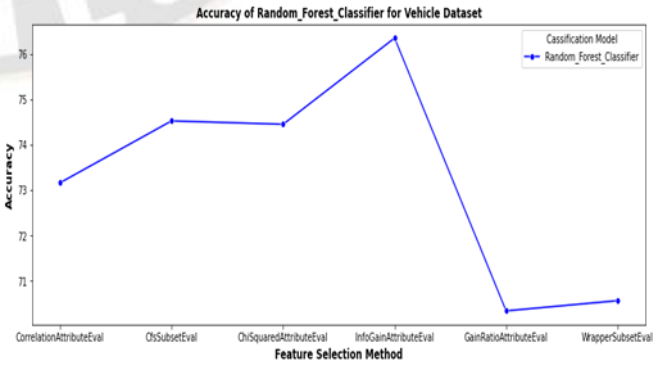

Figure 4 Accuracy of Random Forest Classifier for Vehicle Dataset

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10s
DOI: https://doi.org/10.17762/ijritcc.v11i10s.7643
***Article Received:** 01 June 2023 **Revised:** 30 July 2023 **Accepted:** 12 August 2023*

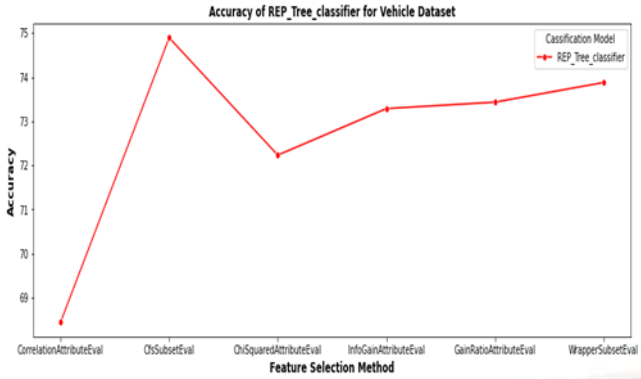

Figure 5 Accuracy of REP Classifier for Vehicle Dataset

For Vehicle dataset, Fig 3 shows that InfoGainAttributeEval + Ranker attribute selection methods with J48 tree classifier accuracy have better accuracy, Fig 4 shows that InfoGainAttributeEval + Ranker attribute selection methods with random forest classifier method perform better accuracy. Fig 5 shows that CfsSubsetEval + Best First attribute selection methods with REP tree classifier accuracy have better accuracy.

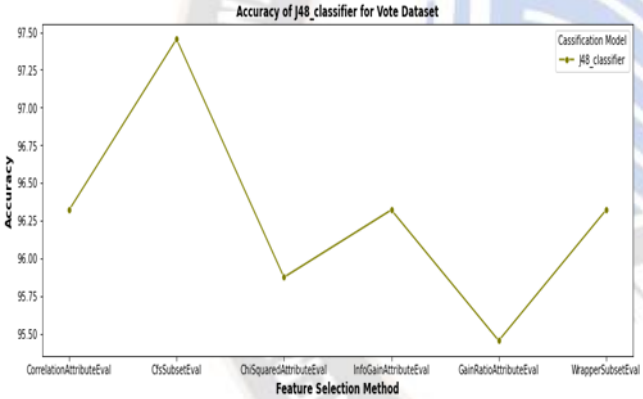

Figure 6 Accuracy of J48 Classifier for Vote Dataset

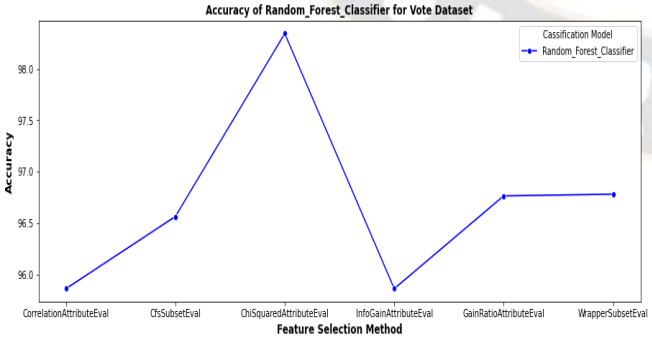

Figure 7 Accuracy of Random Forest Classifier for Vote Dataset

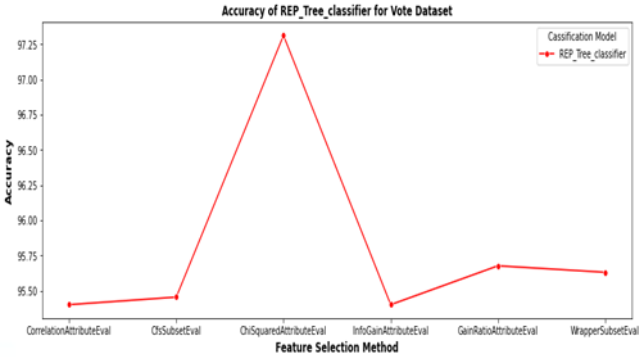

Figure 8 Accuracy of REP Classifier for Vote Dataset

For Vote Dataset Fig 6 shows that CfsSubsetEval + Best First attribute selection methods with J48 tree classifier accuracy have better accuracy, Fig 7 shows that ChiSquaredAttributeEval + Ranker attribute selection methods with random forest classifier method perform better accuracy. Fig 8 shows that InfoGainAttributeEval + Ranker First attribute selection methods with REP tree classifier accuracy have better accuracy.

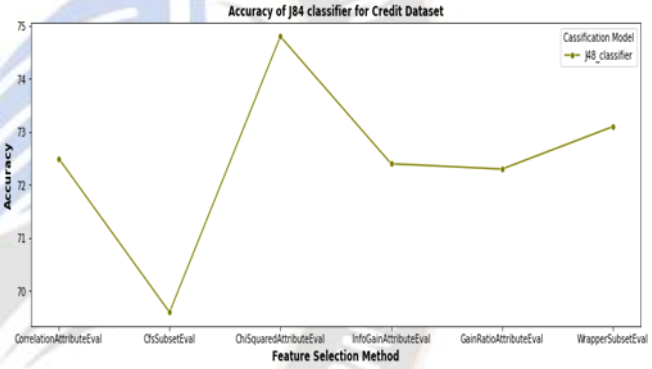

Figure 9 Accuracy of J48 Classifier for Credit Dataset

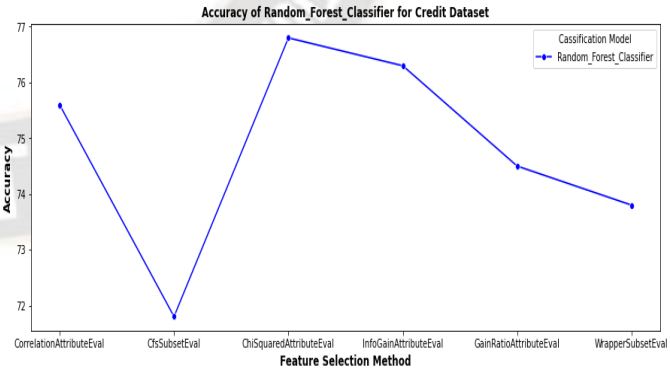

Figure 10 Accuracy of J48 Classifier for Credit Dataset

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10s
DOI: https://doi.org/10.17762/ijritcc.v11i10s.7643
***Article Received:** 01 June 2023 **Revised:** 30 July 2023 **Accepted:** 12 August 2023*

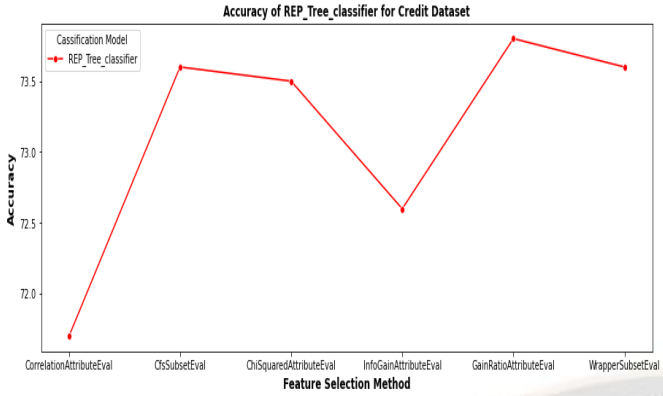

Figure 11 Accuracy of REP Classifier for Credit Dataset

Similarly, For Credit dataset, Fig 9 shows that ChiSquaredAttributeEval + Ranker attribute selection methods with J48 tree classifier accuracy have better accuracy, Fig 10 shows that GainRatioAttributeEval + Ranker attribute selection methods with random forest classifier method perform better accuracy. Fig 11 shows that CfsSubsetEval + Best First attribute selection methods with REP tree classifier accuracy have better accuracy

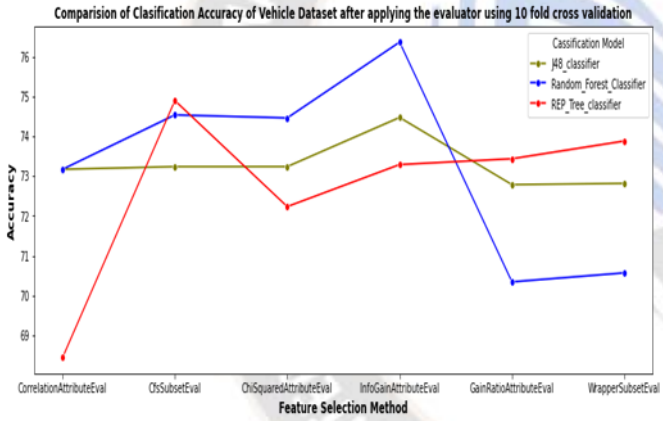

Figure 12 Comparison of Classification Accuracy of Vehicle Dataset

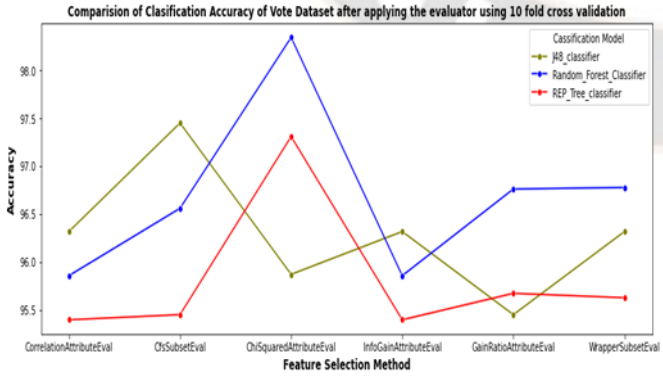

Figure 13 Comparison of Classification Accuracy of Vote Dataset

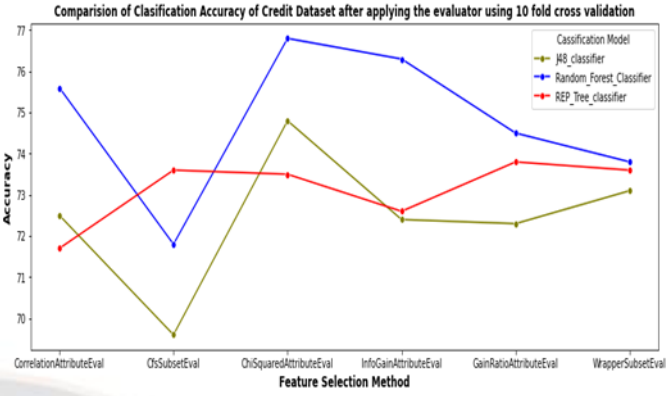

Figure 14 Comparison of Classification Accuracy of Credit Dataset

For the Vehicle dataset, Fig 12 shows that InfoGainAttributeEval + Ranker attribute selection methods with Random Forest classifier accuracy (76.3593) are superior to all other methods. For the Vote dataset, Fig 13 shows that ChiSquaredAttributeEval + Ranker attribute selection methods with Random Forest classifier accuracy (98.345) is greater as compared to other methods. Similarly, For Credit dataset, Fig 14 shows that ChiSquaredAttributeEval + Ranker attribute selection methods with Random Forest classifier accuracy (76.8) is superior to all other methods. In overall comparative analysis, we observed that the feature selection method InfoGainAttributeEval + Ranker with Random Forest classifier perform better result for credit data set and the ChiSquaredAttributeEval + Ranker with Random Forest classifier perform better result for both Vote and Credit datasets, due to this observation the result shows that finally for effective and efficient evaluation the ChiSquaredAttributeEval + Ranker with Random Forest classifier can be used for real time problem.

V. CONCLUSION AND FEATURE WORK

In this paper comparative analysis of feature selection methods with tree based classification has been done on different datasets. In this comparative analysis, we used three datasets to evaluate the performance of six different feature selection methods and their impact on decision tree classifiers using 10-fold cross-validation. The decision tree measures the effectiveness of selected features. Different searching methods are used with each feature section method for selection of optimal subset of features. After the analysis of result, we observed that the feature selection evaluator InfoGainAttributeEval with Ranker searching method perform better result with Random Forest classifier for credit data set and the ChiSquaredAttributeEval evaluator with Ranker search method perform better result with Random Forest classifier for both Vote and Credit datasets. In overall observation, the comparative analysis shows that for effective

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10s
DOI: https://doi.org/10.17762/ijritcc.v11i10s.7643
***Article Received:** 01 June 2023 **Revised:** 30 July 2023 **Accepted:** 12 August 2023*

and efficient evaluation, feature selection method ChiSquaredAttributeEval with Ranker search with Random Forest classifier can be applied for different real datasets in different domains of application.

REFERENCES

[1] Liu Huan and Motoda Hiroshi, "Computational Methods of Feature Selection," Comput. Methods Featur. Sel., vol. 16, pp. 257–274, 2007.

[2] 2. M. A. Hall, "Correlation-based Feature Selection for Machine Learning," no. April, 1999.

[3] R. Kohavi and G. H. John, "Wrappers for feature subset selection," Artif. Intell., vol. 97, no. 1–2, pp. 273–324, 1997, doi: 10.1016/s0004-3702(97)00043-x.

[4] I. Iguyon and A. Elisseeff, "An introduction to variable and feature selection," J. Mach. Learn. Res., vol. 3, pp. 1157–1182, 2003.

[5] P. M. Narendra and K. Fukunaga, "A Branch and Bound Algorithm for Feature Subset Selection," IEEE Trans. Comput., vol. C–26, no. 9, pp. 917–922, 1977, doi: 10.1109/TC.1977.1674939.

[6] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," Comput. Electr. Eng., vol. 40, no. 1, pp. 16–28, 2014, doi: 10.1016/j.compeleceng.2013.11.024.

[7] G. D.E., "Genetic algorithms in search, optimization, and machine learning," Mach. Learn. Reading, Mass, Addison-Wesley Pub. Co, vol. 19, no. SUPPL. 2, pp. 117–119, 1998.

[8] P. Pudil, J. Novovičová, and J. Kittler, "Floating search methods in feature selection," Pattern Recognit. Lett., vol. 15, no. 11, pp. 1119–1125, 1994, doi: 10.1016/0167-8655(94)90127-9.

[9] C. Radhiya Devi, S. K. Jayanthi. (2023). DCNMAF: Dilated Convolution Neural Network Model with Mixed Activation Functions for Image De-Noising. International Journal of Intelligent Systems and Applications in Engineering, 11(4s), 552–557. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/2725

[10] J. Reunanen, "Overfitting in making comparisons between variable selection methods," J. Mach. Learn. Res., vol. 3, pp. 1371–1382, 2003.

[11] A. E. Isabelle Guyon, "An Introduction to Variable and Feature Selection," Procedia Comput. Sci., vol. 94, pp. 465–472, 2016.

[12] P. (Institute for the S. of L. and E. Langley, "Selection of Relevant Features in Machine Learning," Proc. AAAI Fall Symp. Relev., pp. 140–144, 1994.

[13] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," Artif. Intell., vol. 97, no. 1–2, pp. 245–271, 1997, doi: 10.1016/s0004-3702(97)00063-5.

[14] I. M. El-hasnony, H. M. El Bakry, and A. A. Saleh, "Comparative Study among Data Reduction Techniques over Classification Accuracy," vol. 122, no. 2, pp. 8–15, 2015.

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

[16] N. Kwak and C. H. Choi, "Input feature selection for classification problems," IEEE Trans. Neural Networks, vol. 13, no. 1, pp. 143–159, 2002, doi: 10.1109/72.977291.

[17] S. Feedback and A. This, "Feature Selection and Classification Methods for Decision Making : A Comparative Analysis," no. 63, 2015.

[18] A. G. Karegowda, A. S. Manjunath, G. Ratio, and C. F. Evaluation, "Comparative study of Attribute Selection Using Gain Ratio," Int. J. Inf. Technol. Knowl. Knowl. Manag., vol. 2, no. 2, pp. 271–277, 2010, Online]. Available: https://pdfs.semanticscholar.org/3555/1bc9ec8b6ee3c97c524f9c9ceee798c2026e.pdf%0Ahttp://csjournals.com/IJITKM/PDF 3-1/19.pdf.

[19] Z. Zhao, L. Wang, H. Liu, and J. Ye, "On similarity preserving feature selection," IEEE Trans. Knowl. Data Eng., vol. 25, no. 3, pp. 619–632, 2013, doi: 10.1109/TKDE.2011.222.

[20] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy," IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, no. 8, pp. 1226–1238, 2005, doi: 10.1109/TPAMI.2005.159.

[21] O. Osanaiye, H. Cai, K. K. R. Choo, A. Dehghantanha, Z. Xu, and M. Dlodlo, "Ensemble-based multi-filter feature selection method for DDoS detection in cloud computing," Eurasip J. Wirel. Commun. Netw., vol. 2016, no. 1, 2016, doi: 10.1186/s13638-016-0623-3.

[22] E. Ileberi, Y. Sun, and Z. Wang, "A machine learning based credit card fraud detection using the GA algorithm for feature selection," J. Big Data, vol. 9, no. 1, 2022, doi: 10.1186/s40537-022-00573-8.

[23] bin Saion, M. P. . (2021). Simulating Leakage Impact on Steel Industrial System Functionality. International Journal of New Practices in Management and Engineering, 10(03), 12–15. https://doi.org/10.17762/ijnpme.v10i03.129

[24] M. Dash and H. Liu, "Feature selection for classification," Intell. Data Anal., vol. 1, no. 3, pp. 131–156, 1997, doi: 10.3233/IDA-1997-1302.

[25] A. M. M. J. AG Karegowda, "Comparative study of attribute selection using gain ratio and correlation based feature selection," Int J Inf. Technol Knowl Manag, vol. 2, no. 2, pp. 271–277, 2010.

[26] H. Liu, X. Wu, and S. Zhang, "A new supervised feature selection method for pattern classification," Comput Intell, vol. 30, no. 2, pp. 342–361, 2014, doi: 10.1111/j.1467-8640.2012.00465.x.

[27] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," Bioinformatics, vol. 23, no. 19, pp. 2507–2517, 2007, doi: 10.1093/bioinformatics/btm344.

[28] R. Martín, R. Aler, and I. M. Galván, "A filter attribute selection method based on local reliable information," Appl. Intell. 2017 481, vol. 48, no. 1, pp. 35–45, Jun. 2017, doi: 10.1007/S10489-017-0959-3.

[29] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an

update," ACM SIGKDD Explor. Newsl, vol. 11, no. 1, pp. 10–18, Nov. 2009, doi: 10.1145/1656274.1656278.

[30] Anthony Thompson, Anthony Walker, Luis Pérez , Luis Gonzalez, Andrés González. Machine Learning-based Recommender Systems for Educational Resources. Kuwait Journal of Machine Learning, 2(2). Retrieved from http://kuwaitjournals.com/index.php/kjml/article/view/181

[31] A. G. Karegowda, A. S. Manjunath, G. Ratio, and C. F. Evaluation, "Comparative study of Attribute Selection Using Gain Ratio," Int. J. Inf. Technol. Knowl. Knowl. Manag., vol. 2, no. 2, pp. 271–277, 2010.

[32] Pekka Koskinen, Pieter van der Meer, Michael Steiner, Thomas Keller, Marco Bianchi. Automated Feedback Systems for Programming Assignments using Machine Learning. Kuwait Journal of Machine Learning, 2(2). Retrieved from http://kuwaitjournals.com/index.php/kjml/article/view/190

[33] A. G. Karegowda and M. A. Jayaram, "Cascading GA & CFS for feature subset selection in medical data mining," 2009 IEEE Int. Adv. Comput. Conf. IACC 2009, pp. 1428–1431, 2009, doi: 10.1109/IADCC.2009.4809226.

[34] https://archive.ics.uci.edu/ml/datasets/Statlog+(Vehicle+Silhouettes)

[35] https://archive.ics.uci.edu/ml/datasets/congressional+voting+records

[36] https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)