An Approach for Optimal Feature Selection in Machine Learning using Global Sensitivity Analysis

G.Saranya[1]

Research Scholar
Dept. of Computer Science and Engineering
Sathyabama Institute of Science and Technology
Tamil Nadu, India

Dr.A.Pravin[2]

Dept. of Computer Science and Engineering
Sathyabama Institute of Science and Technology
Tamil Nadu, India

Abstract—The classification application is an important procedure for selecting the feature. The classification is mainly based on the features extracted from the object. You can select the best feature using the following three methods: wrapper selection, filter and embedded procedure. All three practices have been implemented by single or combined two approaches. As a result, there is no important feature in the classification process. This problem is solved by the proposed integrated global analysis of sensitivity. Each feature is selected in a classification based on the sensitivity of the feature and the correlation from the target vector in this integrated sensitivity and correlation approach. Likewise, the GSA approach uses a variety of filtering techniques for ranking attributes and optimization using particle swarm technique. Then, the optimum attributes are trained and tested using the Random Forest Classifier grid search via MATLAB software. In comparison to the existing method, wrapper-based selection, the performance of our integrated model is measured using sensitivity, specificity and accuracy. The experimental results of our proposed approach outweigh the sensitivities by 93.72%, 94.74% and the accuracy of 89.921% and 90% where, wrapper selection approach as sensitivity by 89.83% and the accuracy of 93%.

Keywords—Feature selection; feature sensitivity; feature correlation; global sensitivity analysis; classification

I. INTRODUCTION

In a period of large amounts of large amounts of data, social media, health care, bioinformatics, online training and other media are omnipresent. The machine learning methodology depends greatly on useful basic data from a large data pool. Redundant information degrades learning process performance. It is a critical task to determine useful data and remove redundant information. The process for discovering knowledge is used to gather promising information in large data pools. Pre-processing data steps are sub-processed, such as data cleaning, data inclusion, processing and information reduction. A feature is a quantifiable characteristic of a particular process. Feature selection (FS) is the process of selection from a certain dataset for the most important features. In many cases, FS can further enhance the performance of a learning model [1]. A number of features can be used to classify many machines. Real world data has numerous unimportant, superfluous and noisy features. Deleting these characteristics by FS decreases storing and computer costs while preventing important information loss or trying to learn performance degradation. The general approach to FS is shown in Fig. 1.

FS brings numerous benefits as the system's classification increases predictability, knowledge, usability and broad capacity. It also reduces computer systems' complexity and storage, offers a rapid and effective method for knowledge discovery, and plays a critical role. [2] In literature, the majority of FS methods can be classified as wrappers, filters, embedded or hybrid as illustrated in Fig. 2.

Wrapper methods making wise the quality of the selected features by predicting a predefined, learning algorithm. A traditional method takes two different steps, as shown in Fig. 3, according to a particular learning algorithm. (1) a sub-set of features will be searched and (2) the selected feature will be evaluated. [3] Wrapper repeats (1) and until certain stop criteria have been met (2). The search feature set first generates an inferior set of features, but the algorithms studied are then used as black boxes for quality evaluation. For example, the required number of features or the enhanced learning efficiency is achieved iteratively. When a selected function is returned the sub-set of functions provides maximum performance.
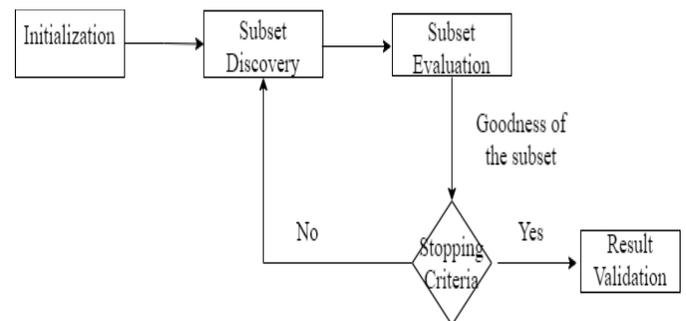

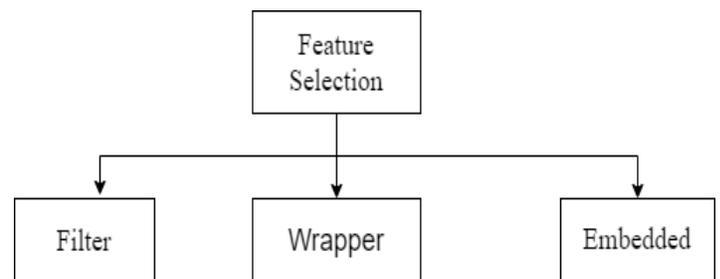Fig. 1. Basic Feature Selection Process.

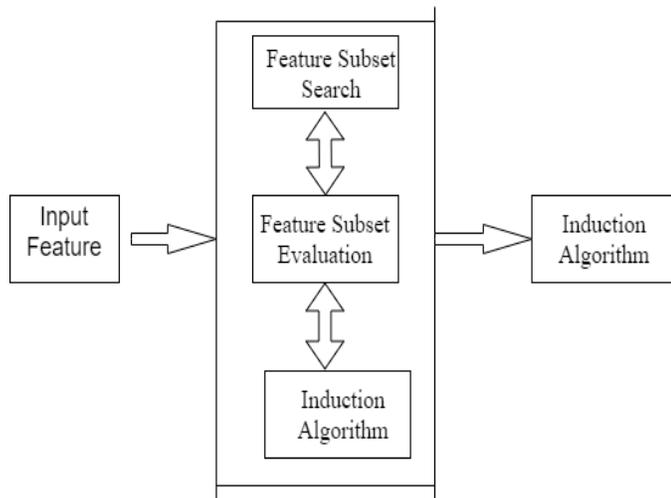Fig. 2. Classification of FS Approaches.

Fig. 3.   Wrapper Feature Selection Working Principle.

In the majority of investigators, the selection of supervised features uses the filter evaluation framework [4]. No learning algorithms are available for the filters. You rely on data properties to measure the importance of its features. In general, filter methods are computationally effective rather than wrapping. However, because the feature selection phase is not governed by any certain algorithm, the selected features cannot be optimal for algorithms for the study. There are two steps to a typical filter method. In the first step, the importance of features is classified by certain criteria for feature assessment. The evaluation process may be one-size-fits-all or multivariate. Low quality features are filtered out in the second phase of the typical filter method. Examples are filter methods [5-9]. A typical FS filter technology diagram is shown in Fig. 4.
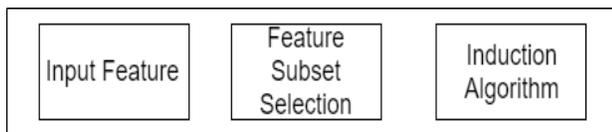

Fig. 4.   Filter based Feature Selection Method.

The last type of feature selection is called embedded methods in Fig 4, a compromise that includes the selection of functions for model learning between the filtering and the wrapping methods [10]. Therefore, these approaches are worthy of wrappers and filter (1) because they need to interact with the classification algorithm; and (2) because they do not require an iterative evaluation of the functional sets. The major integrated approaches are the regulation model that can be adapted to the study model by minimizing fitting mistakes and forcing less coefficients (or exact zero). The official outcome will then be returned, both the regularization model and the chosen functions. In this article we focus on new FS technologies based on a global sensitivity approach.

The paper is organized accordingly: Section II covers the existing related works and its shortcomings. Section III details the work of the proposed methodology. Section IV describes implementation and the results are discussed. Section V summarizes its performance to conclude this document. The future improvement work was suggested in Section VI.

II.   RELATED WORKS

A mixture of several methodologies can be considered hybrid techniques to select features (e.g., filters, wrapper, and embedded). The main goal of many traditional approaches for selection is to solve problems of instability and interruption. Hybrid methods are examples of [11-12]. A small data disturbance, for example, can lead to completely different selection results in small, high-dimensional data. The results are consistent and therefore the integrity of the selected functions is managed to improve by combining several selected subassemblies from different methods.

Table I shows the comparative analysis of various existing feature selection techniques in the classification cardiovascular diseases. The support vector machine was preferred as classifier by most algorithms. It is due to the classification of the binary class. Due to the hierarchical arrangement of data only few works were suggested by the random forest classifier. The wrapper selection technique allows the three filter techniques to be highly precisely designed. This is due to the attribute's selection process. The attributes are selected according to the classifier's performance. On this basis, the current Cleveland data set classification technology is selected for the wrapper feature. You can use the grey-wolf optimization algorithm [13] to choose a feature from the dataset here. The selected characteristics were then given training and variable classified. It can achieve 89.83% higher grading rates. However, the wrapper selection algorithms have the following problems.

The wrapper method depends entirely on the grading option. With a different classification, the result can vary.

The classification efficiency determines the selection of fitness and variables.

These disadvantages have been overcome through the proposed global classification based on sensitivity analysis. The following Section III provides a brief explanation of the proposed approach.

TABLE I.   COMPARATIVE STUDY OF DIFFERENT METHODS OF EXISTING FEATURE SELECTION

Author	Type	Technique	observations
Subanya and Rajalakshmi [13]	Wrapper selection	Artificial bee colony algorithm on heart diseases dataset SVM classification is performed Only 7 features were used.	• Fitness is not defined. • The results do not verify their identification nature • accuracy is the only measure.
Shah et al. [14]	Wrapper selection	Selection of features: mean fishing score and accuracy Reduction of feature: principal analysis component. Radial base kernel-based classification Vector support machine support	• Computational time high due to multiple feature extraction. • Cleveland datasets have minimal accuracy when compared to other datasets such as Hungary and Switzerland.
Chiroma et al. [15]	Wrapper selections	Rule based algorithm like Prism is used for feature selection and classification Wrapper algorithms like decision tree, LibSVM, K-nearest neighbor were used. Wrapper algorithm utilized forward selection and backward elimination algorithm	• Cleveland's f-score is a minimum of the other dataset. • As with prism, the algorithm of the Cleveland wrapper can also produce better results.
Moorthy and Gandhi [16]	Wrapper selection	Combination of ANOVA and whale optimization is used. Classification: naïve bayes, SVM and K-NN	The whale optimization algorithm for Cleveland data sets in any classification type can achieve greater precision. Compared to others, SVM produced Cleveland's best results in classification. However, it used all the classification attributes.
Wang and Li [17]	Wrapper selection	Immune system artificial	The best sonar and cardiac statolog result are produced. It can be as accurate as the Cleveland data set approach.
Saqalin et al. [18]	Filter selection	Selection of the feature: fisher score, selection forward and selection backward. Reduction feature: coefficient of mathematical correlation Classification: SVM based on radial basis	Cleveland's 81 percent data set is less accurate. The Switzerland heart disease dataset produced the best result. Less feature reduces the performance of the classification.
Garate Escamila et al. [19]	Filter selection	Chi-square and Principal Component analysis is used for feature selection. Classifier: six machine learning algorithms is used	• Random forest produced the best classification result. • Processing time is high for more attributes.
Gupta et al. [20]	Filter selection	Correlation and Squared correlation are used for feature selection. Random forest classifier is best among many classifiers	Able to achieve high classification rate. The number of features is high.
Ayon et al. [21]	Filter selection	Feature selection: Correlation Classifier: logistic regression, and many machine learning algorithms	Five-fold SVM based classification produced the best result. Real time features require modification
Muhammad et al. [22]	Ensemble Selection	Feature selection: LASSO, Relief, mRMR Different classifiers.	Proposal is made based on the analysis. Optimization techniques can improve the results further.

III. PROPOSED METHODOLOGY

This work introduces efficient approaches to select features such as integrated sensitivity and correlation and a global hybrid feature sensitivity analysis. We're using the data set for cardiac disease in Cleveland [23] here. It consists of 76 variables, 14 of which were only selected for use. Of the 14 attributes, 13 are predictors, and the final attributes are the target. In studies, 270 cases, 120 of which were categorized as CHD patients and 150 cases as CHD-free patients, were considered for the elimination of missing-value cases [14]. The characteristics and range of values are explained in Table II.

TABLE II.   CLEVELAND HEART DISEASE DATASET [23]

Features	Description	Ranges
age	Age (in years)	29-77
sex	Gender	1: male; 0: female
Cp	chest pain type	0-3
Trestbps	resting blood pressure	94-200
Chol	serum cholesterol in mg/dl	126-564
Fbs	Fasting blood sugar	0-1
Restecg	resting electrocardiographic results	0-2
thalach	maximum heart rate achieved	71-202
Exang	exercise induced angina	0-1
Oldpeak	ST depression induced by exercise relative to rest	0-6.2
Slope	the slope of the peak exercise ST segment	0-2
ca	number of major vessels (0-3) colored by flourosopy	0-4
thal	3 = normal; 6 = fixed defect; 7 = reversable defect	0-3

A. Feature Sensitivity

The first stage uses the sensitiveness approach to identify the sensitivity to output variations in a particular model for each input factor. The results from the feature sensitivity analysis are highly dependent on the factors to be carefully selected. We have here classified their significance in determining the CHD risk. Ranks have been determined using functional sensitivity in a learned classification algorithm. After removing the least preferred characteristics, design was gradually trained according to these rankings. This phase continued until compared with the previous one the model performance deteriorated. In this approach, we examine the model, in order to analyze the differences between the characteristics of the development of the learning model.

The sensitivity of the ith feature Sen (M, m_i) is determined by a condition that differs between the original and the deformed data set by adding very little noise (known as μ) in the developed model.

$$\text{Sen } (M, m_i) = \frac{1}{N} \sum_{\forall k} | \text{ RFOutput}_k \left(M_{(m_i + \delta)} \right) - \text{RFOutput}_k(M) |$$

(1)

In the RFOutputk(M) and RFOutputk(M(mi + δ), the inputs k are the output, with the original input data set M, and then the result with a noisy input (X(xi + δ)) is the very small noise δ to i^{th}. All sensitivities were measured individually with a single sensitivity. μ was randomly selected in the range [a1, 0.0010].:

B. Feature Correlation

We analyze the characteristics of model prediction outcomes and evaluate them. When changes were affected by features in the input for the preview performance, features were deemed correlated. This means that the value of the property is increased when training the model if a feature improves its severity. In addition, the relating characteristics can be compared if the size of the increase greatly affects the other features. Selection of features Correlation sees the class and value-based correlation of the subfunctions as an ideal set of characteristics:

$$\text{CFS} = \max_{s_i} \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)r_f}}$$

(2)

Where (r_cf) $\perp^-$ the average correlation of all the features is equal to rf‾ the average correlation of all features-class.

In this study, we analyzed features for category relationships and evaluated if they were correlated with the results of classification predictions mainly during the feature correlation analysis stage. If any of the features affects the possibility that the correlation will contribute to the output, the features were considered to be correlated.

Fig. 5 shows the flow diagram for our proposed integrated sensitivity model and correlation. The flowchart begins with the data pre-processing where missing data is processed using a data imputation method followed by min-max standardization. Any negligible input features can reduce the output of the classifier. It is therefore very difficult to select from a collection of features for the prediction mission an exact and rigorous set of attributes. Feature selection is made through the combination of feature sensitivity and functional correlation in the presented design. Each approach will evaluate the rating of features and then measure the value of the response variable by using the amount from both approaches. Functions are selected first in order to increase 1 to 13, and then in the second scenario, in order to reduce 13 to 1. This can be used to calculate and check the optimal feature subset. Random forest models are generated after the input rating with different number of features in order to estimate heart disease. A novel integrated feature selection result is linked in comparison with existing classification models like naive Bayes, decision tree, regression analysis and support vector machine.

The following is a pseudocode for feature selection based on sensitivity analysis:

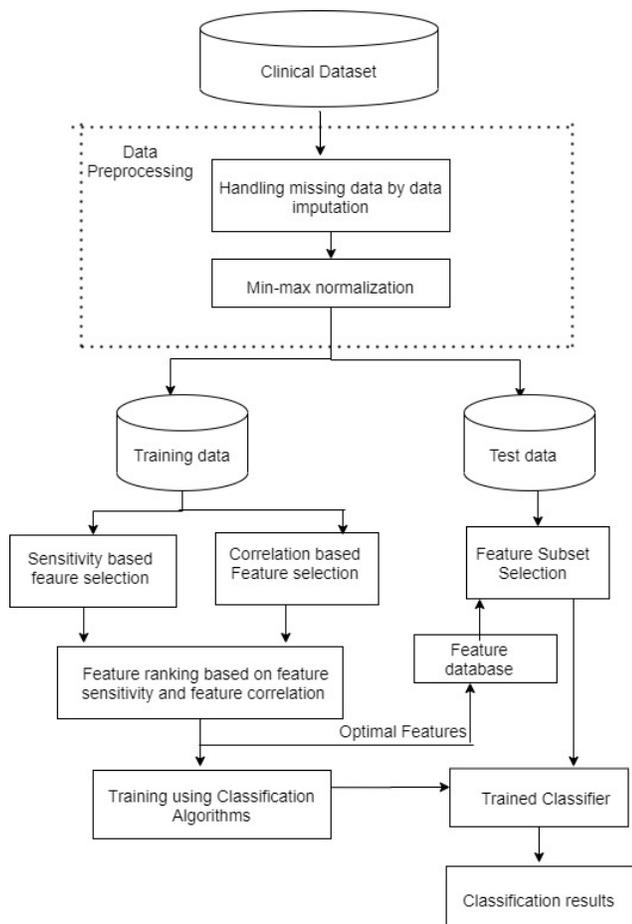| Input: X = x1, x2, x3, …xi.... xn /∗ features of Cleveland dataset |
|---|
| #Choose the feature subset based on sensitivities |
| #Assess all feature sensitivities |
| # Rank the characteristics according to their sensitivities |
| #Add the features according to their rank to the feature subset |
| Output: *X*s /∗ Chosen feature sub-set */ |

Fig. 5.   Flowchart of Integrated Feature Sensitivity and Correlation Approach.

C. Global Sensitivity analysis - Particle Swarm optimization Feature Selection

This proposes an optimized filter technique for the classification process feature selection as shown in Fig. 6. There have been two phases in which to select classification attributes. The very first stage is based on a global analysis of feature sensitivity. In the second stage, a wrapper optimization determines the predominate attribute from the first phase. Due to this multi-stage functions selection, the framework developed can be used to apply every type of machine learning application. The purpose of the approach proposed is to attain the following objectives.

1) The method proposed does not depend on the first phase ranking of the selection of features.

2) Larger data sets can offer greater precision.

The importance of the attribute is determined by three individual filtering approaches:

1) The coefficient of correlation of the input vector to the objective output is calculated by Pearson. The correlation values are -1 and 1. The classification depends on the negative correlation approach. The attribute is listed below and the negatively correlated attribute is higher. Each ranking of attributes is performed on this basis.
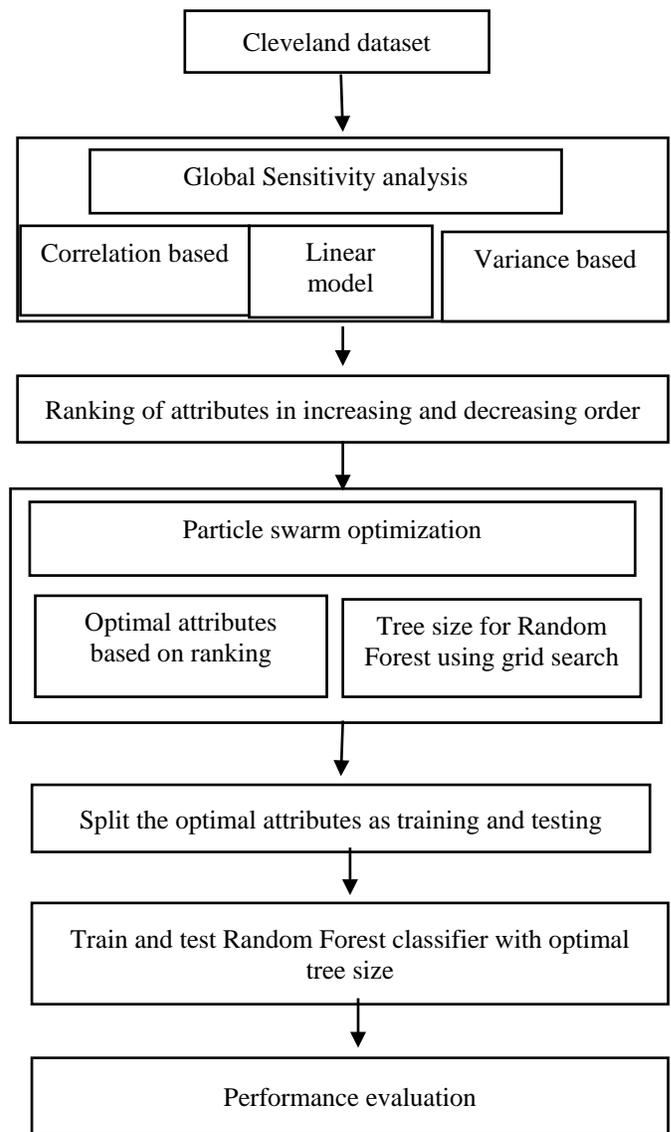

Fig. 6.   Flowchart of GSA-PSO Feature Selection.

2) Linear model fitting- The value of the attribute is determined by the minimum average vector input error. Input attributes X1 to Xn and destination Y are included in this data set.

3) Variance based- The variance between the input and the target vector is calculated in order to determine its importance. Two steps for determining the difference. First, it calculates the medium of the attribute. The significant difference is also used to calculate the difference between the single attribute and its mean value.

D. Classification Methods

1) Random forest: The Random Forest (RF) as a supervised method of learning has been introduced recently to engineering practice [23–25]. This RF procedure combines two powerful ML techniques, bootstraps [26] and a random subspace [27]. It is therefore excellent to generalize this technique, as it adds the results of several decision-making

bodies, but computer cost is substantial. N arbitrary samples and trainings are derived using the bagging technique in this algorithm. Bootstrap sets are used to decide trees. Each node tests a feature and the leaf nodes are the output labels. The solution is achieved by combining all of the outputs [28-29], as follows: The solution is:

$$m = \frac{1}{n_{\text{tree}}} \sum_{i=1}^{n_{\text{tree}}} m_i(x) \qquad (3)$$

Where y = the mean output of the total ntree amount; m_i(x) = the prediction of the individual tree for the vector input x.

Fig. 10 illustrates the RF model comparison with other learning models. Fig. 11 shows the RF model comparison to the Proposed RF FSFC.

Pseudocode for feature selection based on GSA-PSO feature selection is given below:

Inputs: Cleveland dataset with thirteen attributes (Mn) and two classes (Pn)
Begin GSA
CR=corr(Mn,Pn) /* GSA based on Correlation*/
LM=fitlm(Mn,Pn) /*GSA based on Linear fit Model*/
VM=Var (Mn,Pn) /* GSA based on Variance-based*/
End GSA
#Ranking of attributes in increasing and decreasing order
Begin PSO /*Particle Swarm optimization*/
#Optimal attributes based on ranking
#Find Tree size for random forest using Grid search
End PSO
#Simulate the model with train and test/* Split the optimal attributes as training and testing*/
#Evaluate the model Performance
Output: Performance evaluation, Per. /* Performance evaluation of the classifier with optimal tree size*/

E. Optimization of the Feature and Size of the Tree by Particle Swarm

The optimal attribute and tree size is determined for the random forest classification via an optimization approach. The best solution for the problem is to solve the fitness function. The fitness focus is to prevent the rate of errors in the random forest classification. The fitness function is given with the following equation 5.

$$FitnessFunc_{GSA-PSO} = \\ minimum(error\ rate\ of\ RandomForest) \qquad (4)$$

This is achieved by finding the optimum attribute through the grid search algorithm based on a global sensitivity analysis and the best tree size between 10 and 130. By minimizing the classificatory error rate on the search algorithm, the optimum tree can also be determined. The common fitness function for optimizing the algorithm is therefore used.

IV. Implementation, Results and Discussions

With the MATLAB Software R2018a, the proposed method is presented in Windows 10. The selection of characteristics based on integrated sensitivity and correlation is computed and the selection of characteristics calculated using the Global Particle Swarm Optimization Sensitivity analysis. And the approaches are comparted on the basis of accuracy, sensitivity, specificity and time. The Table III shows an integrated selection of sensitivities and correlation functions. Therefore, the sensitivity of properties is determined by Equation 1 and the correlation between properties is calculated by Equation 2. The higher the sensitive value and the lower the sensitive one. And the less correlated feature is higher. In the case of the calculation of a total function range, features of the same rank will be given priority.

By combining the ranking of these techniques and identifying the importance of each attribute in classifying the data set, the integrated feature selection procedure by feature sensitivity and correlation is performed. Table III shows the general classification of these two methods for each attribute. Likewise, in order for the importance of attributes to be determined during the grading process, there was a combination of rankings with Pearson's correlation and variance-based tests as shown in Table IV. There have been sensitivity analyses.

The classified attributes were arranged in two ways in an increasing and decreasing order to determine the optimized attributes for the classification process. The ranking of lower to higher attributes is indicated by the increasing order. Lower order indicates the ranking of the higher to lower attributes.

The Cleveland dataset is used for random classification based on the above optimal attributes. The classification system is trained to increase order formats with the optimal attributes of 1 to 13 percent and 70 percent of the data. The remaining 30 percent data is then used to test and evaluate trained classifiers.

The greater ranking of an integrated method that allows patients with 93.72% sensitivities effectively to identify cardiac diseases compared with the control patient with 83.28% specificity. The overall precision of the increased order rating is 89.921%.

TABLE III.   RANKING OF ATTRIBUTES WITH INTEGRATED FEATURE SENSITIVITY AND CORRELATION

S.no	Attribute_name	Integrated Sensitivity and Correlation		Ranking			Overall Ranking Score	
		Sensitivity	Correlation	S	C	Total Rank	Increasing order	Decreasing order
1	age	0.024	0.2254	6	8	14	6	8
2	sex	0.048	0.2809	5	9	14	7	7
3	Cp	0.003	-0.4338	13	1	14	4	10
4	Trestbps	0.011	0.1449	9	7	16	10	4
5	Chol	0.005	0.0852	11	6	17	11	3
6	Fbs	0.1	0.028	1	5	6	1	13
7	Restecg	0.049	-0.1372	4	4	8	2	12
8	thalach	0.004	-0.4217	12	2	14	5	9
9	Exang	0.082	0.4368	2	13	15	9	5
10	Oldpeak	0.073	0.4307	3	12	15	8	6
11	Slope	0.01	-0.3459	10	3	13	3	11
12	ca	0.012	0.3917	8	11	19	13	1
13	thal	0.013	0.344	7	10	17	12	2

TABLE IV.   RANKING OF ATTRIBUTES USING GSA APPROACH

S.no	Attribute_name	Sensitivity factors			Ranking			Total Rank	Overall Ranking Score	
		Variance	Pearson	Significance	V	P	L		Increasing Order	Decreasing Order
1	age	82.21233	0.2254	0.76113	4	8	12	24	8	6
2	sex	0.216449	0.2809	4.2449e-05	12	9	3	24	9	5
3	Cp	1.061617	-0.4338	8.4015e-07	6	1	1	8	1	13
4	Trestbps	306.5713	0.1449	0.11441	3	7	9	19	5	9
5	Chol	2677.560	0.0852	0.40255	1	6	11	18	3	11
6	Fbs	0.126458	0.0280	0.77112	13	5	13	31	13	1
7	Restecg	0.275615	-0.1372	0.21282	10	4	10	24	10	4
8	thalach	522.9148	-0.4217	0.0079882	2	2	5	9	2	12
9	Exang	0.219978	0.4368	0.0053868	11	13	6	30	12	2
10	Oldpeak	1.343646	0.4307	0.010847	5	12	8	25	11	3
11	Slope	0.378481	-0.3459	0.063453	8	3	7	18	4	10
12	ca	1.042272	0.3917	6.2486e-06	7	11	2	20	6	8
13	thal	0.373645	0.3440	0.00095231	9	10	4	23	7	7

The reduction of ranking classifications is also trained, and all its attributes are tested using the random classification using 70% of data and 30% of data. In comparison with normal patients in the lower classifications, it can also efficiently detect cardiac patients. Its total precision, however, is 78.32%.

Table V and Fig. 7 shows the comparison between the two approaches proposed by the use of performance assessment in increased and decreasing order.

TABLE V.   COMPARISON OF THE INTEGRATED FSFC APPROACH TO PERFORMANCE ASSESSMENT

Method	Increased order ranking (%)	Decreased order ranking (%)
Accuracy	89.921	78.32
Sensitivity	93.72	80.23
Specificity	83.28	78.47

Performance evaluation comparison of Integrated FSFC approach

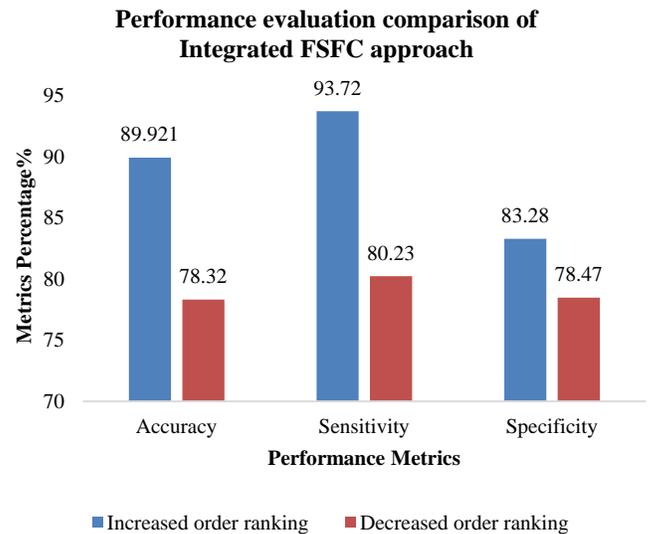

Fig. 7.   Comparison of the Integrated FSFC Approach to Performance Assessment.

TABLE VI.   COMPARISON OF GSA APPROACH PERFORMANCE ASSESSMENT

Method	Increased order ranking (%)	Decreased order ranking (%)
Accuracy	90.00	73.33
Sensitivity	94.74	69.23
Specificity	81.82	76.47

TABLE VII.   PERFORMANCE COMPARISON BETWEEN GSA AND EXISTING WRAPPER SELECTION

Method	GSA analysis (%)	Integrated sensitivity and correlation analysis (%)	Existing wrapper selection GWO-SVM
Accuracy	90.00	89.921	89.83
Sensitivity	94.74	93.72	93
Specificity	81.82	83.28	91

Fig. 8.   Performance Evaluation Comparison of GSA Approach.

Fig. 9.   Performance Comparison between Proposed Approaches and Existing Wrapper Selection.

Similarly, with 94.74% sensitivity, 81.82% specialty and the overall accuracy for the increased ranking classifier is 90.00%, the GSA method. And with a sensitivity of 69.23 percent, 76.47 percent are declining and 73.33% are indicated in Table VI and Fig. 8 for the total accuracy of the classifier.

The performance evaluation shows that the higher rankings can effectively classify heart disease by having higher precision compared to decreasing order precision.

This means that in the Cleveland dataset, the classification of patients with normal and heart disease is best determined by the increasing order.

Then, the proposed integrated approach and global sensitivity analysis approach is compared to the existing wrapper selection method with the same Cleveland data set, using a grey wolf optimization support vector machine classifier.

In addition, our approaches, when compared with the existing wrapper selection performance, are shown in Table VII and Fig. 9 compared with the current wrapper 93% above the wrapper selection by precise determination of cardiac diseases with high sensitivity at 93.72% and 94%.
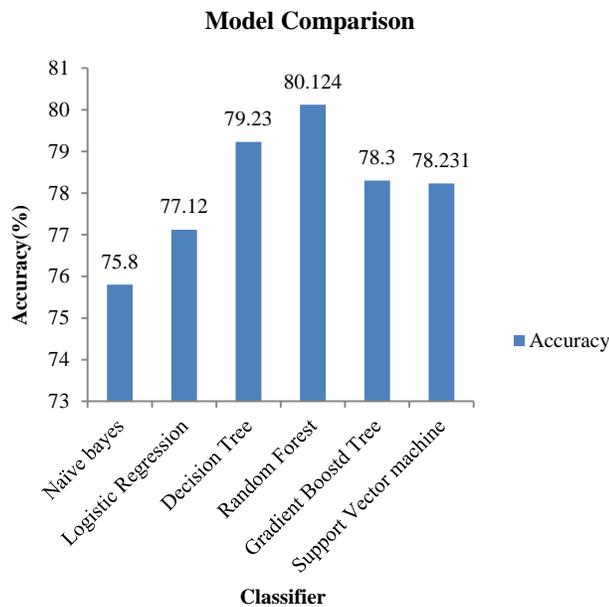
Model Comparison

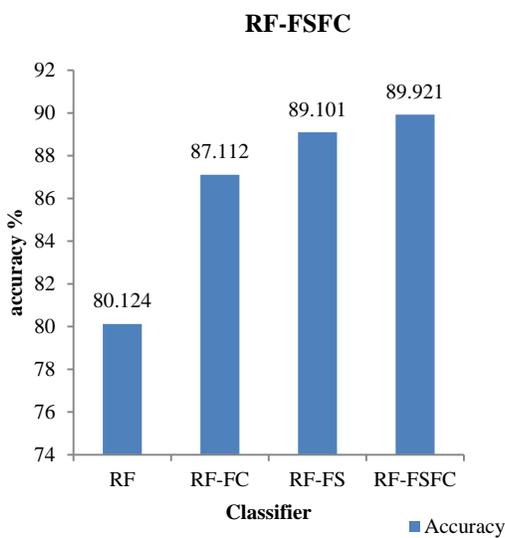Fig. 10. Comparison of RF Model with the Other Learning Models.

RF-FSFC

Fig. 11. Comparison of RF Accuracy based on Sensitivity and Correlation.

V. Conclusion

The purpose of the features is assessed based on an integrated approach to sensitivity and correlation and a global sensitivity analysis based on optimization. The integrated sensitivity and correlation of properties are implemented in two phases as follows:

1) The first stage is to rank each attribute based on feature sensitivity analysis between the vector and the target.

2) The second phase consists of classifying each attribute by a correlation between the vector function and the variable objective.

The global sensitivity analysis is similarly conducted in two phases:

1) The first step is a classification based on different sensitivity analyses of each attribute.

2) The second Phase defines a more sensitive grading attribute based on the optimization of particle swarm.

The proposed approaches assess the ranking of the random forest classifier's trees and the reduced order ranking with optimum dimensions. It then uses its accuracy, sensitivity and specificity to assess its performance. The lower to higher classification produced the best results for heart disease patients. This classification helps to avoid the least classified features.

By comparing the performance with the wrapper selection-based classification, our proposed approaches integrated feature sensitivity and feature correlation outperforms it with the accuracy 89.921% and the sensitivity with 93.72% and also the global sensitivity analysis outperforms the wrapper selection by finding heart diseases accurately with the accuracy 90% and high sensitivity of above 94% as compared to existing of 93%.

The integrated approach to feature sensitivity and correlation and the global approach to sensitivity play a significant role in choosing the best classification feature especially in comparison to each feature selection procedure.

VI. Future Enhancement

With the aim of reducing calculation times for the selection of attributes and improving the performance of the learning model, the results from our suggested approaches can be further strengthened.

References

[1] A Shroff, K. P., & Maheta, H. H. (2015). A comparative study of various feature selection techniques in high-dimensional data set to improve classification accuracy. 2015 International Conference on Computer Communication and Informatics (ICCCI). https://doi.org/10.1109/iccci.2015.7218098.

[2] Kishore, R., & Tripathi, S. (2016). A comparative analysis of enzyme classification approaches using hybrid feature selection technique. 2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT).

[3] El Aboudi, N., & Benhlima, L. (2016). Review on wrapper feature selection approaches. 2016 International Conference on Engineering & MIS (ICEMIS). https://doi.org/10.1109/icemis.2016.7745366.

[4] Yin, P., Mao, N., Zhao, C., Wu, J., Sun, C., Chen, L., & Hong, N. (2018). Comparison of radiomics machine-learning classifiers and feature selection for differentiation of sacral chordoma and sacral giant cell tumour based on 3D computed tomography features. European Radiology, 29(4), 1841-1847.

[5] Saranya, G., & Pravin, A. (2020). Feature selection techniques for disease diagnosis system: A survey. Artificial Intelligence Techniques for Advanced Computing Applications, 249-258. https://doi.org/10.1007/978-981-15-5329-5_24.

[6] M Robnik-Šikonja, I Kononenko, "Theoretical and empirical analysis of ReliefF and RReliefF" - Machine learning, 2003 – Springer.

[7] Feng Yang and K. Z. Mao. 2011. Robust feature selection for microarray data based on multicriterion fusion. IEEE/ACM Trans. Comput. Biol. Bioinform. 8, 4 (2011), 1080–1092.

[8] Lei Shi, Liang Du, and Yi-Dong Shen. 2014. Robust spectral learning for unsupervised feature selection. In ICDM. 977– 982.

[9] Hiromasa Arai, Crystal Maung, Ke Xu, and Haim Schweitzer. 2016. Unsupervised feature selection by heuristic search with provable bounds on suboptimality. In AAAI. 666–672.

[10] Y Guo, J Ji, H Huo, T Fang, D Li, "SIP-FS: a novel feature selection for data representation on Image" 2018 - jivp-eurasipjournals.springeropen.

[11] X Liu, J Tang, "Mass classification in mammograms using selected geometry and texture features, and a new SVM-based feature selection method" - IEEE Systems Journal, 2014 - ieeexplore.ieee.org.

[12] D Singh, AA Gnana, "A novel feature selection method for image classification" 2015.

[13] Subanya, B., & Rajalaxmi, R. R. (2014, February). Feature selection using Artificial Bee Colony for cardiovascular disease classification. In 2014 International Conference on Electronics and Communication Systems (ICECS) (pp. 1-6). IEEE.

[14] Shah, S. M. S., Shah, F. A., Hussain, S. A., & Batool, S. (2020). Support Vector Machines-based Heart Disease Diagnosis using Feature Subset, Wrapping Selection and Extraction Methods. Computers & Electrical Engineering, 84, 106628.

[15] Chiroma, F., Cocea, M., & Liu, H. (2019). Evaluation of rule-based learning and feature selection approaches for classification. OASIcs.

[16] Moorthy, U., & Gandhi, U. D. (2020). A novel optimal feature selection technique for medical data classification using ANOVA based whale optimization. Journal of Ambient Intelligence and Humanized Computing, 1-12.

[17] Wang, Y., & Li, T. (2020). Local feature selection based on artificial immune system for classification. Applied Soft Computing, 87, 105989.

[18] Saqlain, S. M., Sher, M., Shah, F. A., Khan, I., Ashraf, M. U., Awais, M., & Ghani, A. (2019). Fisher score and Matthew's correlation coefficient-based feature subset selection for heart disease diagnosis using support vector machines. Knowledge and Information Systems, 58(1), 139-167.

[19] Gárate-Escamila, A. K., El Hassani, A. H., & Andrès, E. (2020). Classification models for heart disease prediction using feature selection and PCA. Informatics in Medicine Unlocked, 19, 100330.

[20] Gupta, A., Kumar, R., Arora, H. S., & Raman, B. (2019). MIFH: A machine intelligence framework for heart disease diagnosis. IEEE Access, 8, 14659-14674.

[21] Ayon, S. I., Islam, M. M., & Hossain, M. R. (2020). Coronary artery heart disease prediction: a comparative study of computational intelligence techniques. IETE Journal of Research, 1-20.

[22] Muhammad, Y., Tahir, M., Hayat, M., & Chong, K. T. (2020). Early and accurate detection and diagnosis of heart disease using intelligent computational model. Scientific reports, 10(1), 1-17.

[23] Cleveland Clinic Foundation, "Heart Disease Data Set ", Available at: http://archive.ics.uci.edu/ml/datasets/Heart+Disease.

[24] J. Zhou, X. Shi, K. Du, X.Y. Qiu, X.B. Li, H.S. Mitri, Feasibility of random-forest approach for prediction of ground settlements induced by the construction of a shield-driven tunnel, Int. J. Geomech. 17 (2016) 04016129.

[25] Y. Zhou, S.Q. Li, C. Zhou, H.B. Luo, Intelligent approach based on random forest for safety risk prediction of deep foundation pit in subway stations, J. Comput. Civ. Eng. 33 (2019) 05018004.

[26] P. Zhang, R.-P. Chen, H.-N. Wu, Real-time analysis and regulation of EPB shield steering using random forest, Automat. Constr. 106 (2019) 102860.

[27] Saranya, G., & Pravin, A. (2020). A comprehensive study on disease risk predictions in machine learning. International Journal of Electrical and Computer Engineering (IJECE), 10(4), 4217. https://doi.org/10.11591/ijece.v10i4.pp4217-4225.

[28] Hemanth Reddy, K., & Saranya, G. (2020). Prediction of cardiovascular diseases in diabetic patients using machine learning techniques. Artificial Intelligence Techniques for Advanced Computing Applications, 299-305. https://doi.org/10.1007/978-981-15-5329-5_28.

[29] A. Liaw, M. Wiener, Classification and regression by random forest, R News 23 (2002) 18–21.