Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 3892 - 3913
Received 20 January 2021; Accepted 08 February 2021.

Optimal Feature Subset Selection Method for Improving Classification Accuracy of Medical Datasets

C.Sathish Kumar,
Research Scholar, Bharathidasan University, Tiruchirappalli &
Associate Professor, PG & Research Department of Computer Science,
Bishop Heber College (Autonomous), Affiliated to Bharathidasan University,
Tiruchirappalli, Tamilnadu, India.
(Email: satgreen.in@gmail.com)

P.Thangaraju,
Associate Professor, PG & Research Department of Computer Science,
Bishop Heber College (Autonomous), Affiliated to Bharathidasan University,
Tiruchirappalli, Tamilnadu, India.
(Email: trthangaraju@gmail.com)

Abstract

Medicine is indeed one of the sciences in which computer science advancement makes a great deal of progress. The use of medical computers enhances precision and speeds data processing and diagnosis processes. There are now various diagnostic systems assisted by computers and machine-learning algorithms play a key role. More precise and faster systems are needed. The classification is the popular machine-learning job, part of computer-aided diagnostic systems and various packages of medical data analysis software. It is necessary to choose functional set and proper parameters for the classification model to achieve higher classification accuracy. Medical databases also contain wide set of features where many features correlate with others, so reducing the set of features is essential. Most classifiers are structured so that they can learn from the data themselves through a training process because full expert experience is not realistic to evaluate classification parameters.In this paper, in order to improve the accuracy of the classifier with proposed Feature Selection method. Differential evolution optimization is used to find the optimal subset obtainedby the Filter based feature selection method.The performance of the proposed feature selection is evaluated with classifiers like Random Forest Classifier, Gradient Boosting Tree, Artificial Neural Network and Support Vector Machine.

Keywords: Machine Learning, Medical dataset, Feature Selection, Filter based approach, Classifications, Random Forest, Gradient Boosting, Artificial Neural Network, Support Vector Machine.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 3892 - 3913
Received 20 January 2021; Accepted 08 February 2021.

1. INTRODUCTION

Progress in the field of computer science and technology will help in a variety of areas, including medicine. Quick and correct diagnoses in medicine can save a patient's life, so that it is important that good Computer Aided Diagnostic systems (CADs) are available to assist doctors. Classification [1] is typically the most critical aspect in CAD systems.

Classification is one of the tasks of machine learning when machine learning algorithms are utilized to collect some information from a number of data, to look for some patterns and then to determine for themselves based on facts learned. Machine learning algorithms are currently widely used and studied, since they are used in different areas, including medicine, bioinformatics, economics, agriculture and robotics. Classification is an educational activity that is supervised, where the performance is graded, where certain instances belong. Monitored learning is a method in which a decision model is developed to better identify unknown cases based on the model based on a collection of training courses known to the classes. A model of decision searching for training data set models that allow new unknown instances to be classified[2][3].

Medical dataset classification is a challenge since it typically has a large number of features and instances. The need for early and accurate diagnostic therapies to recover a patient is pushed into the search for more specific and faster CAD methods of classification. In a classification task each instance is shown with many numeric or categorical features. The accuracy of classification depends heavily on the selected features, in addition to the classification process, which allows the classification method to distinguish instances from different classes and to find similarities between instances in the same class. It is hard to determine which characteristics characterise an instance so that when gathering data, the normal approach is to define instances with as much functionality as possible and then to decide which ones are relevant. Too many characteristics will lead to the problem of reducing the impacts of key discrepancies and similarities in the decision model, as all possible details will be included [4]. Therefore, the issue of feature selection attracts scientists and is one of the topics of study. The objective of the feature selection methods is to define the minimum feature subset that provides the best ranking. The problem of selection of feature is an exponential one. There are 2n subsets for a set with n features. It means that even for relatively little values of n, exhaustive search is not possible (in reasonable time).

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 3892 - 3913
Received 20 January 2021; Accepted 08 February 2021.

Metaheuristics like swarm intelligence and optimization algorithms can be used for solving problems like this.

2. IMPORTANCE OF FEATURE SELECTION

The precision of the classification system depends not only on the classification algorithm but also on the method of selection. Choosing inappropriate and irrelevant features can confuse the classifier, resulting in incorrect results. The solution is to pick a feature, i.e., to increase classification efficiency and accuracy, the feature selection is important. Select the sub-set of features from the original set by eliminating the irrelevant and redundant features from the original dataset. The feature collection It is also called collection of attributes. Featureselection [5] decreases data set dimensionality, increases accuracy of learning and improves understandability of the outcome. In the collection and elimination of the required feature the two search algorithms for the forward collection and the backward elimination. Feature selection is the three-step search, evaluation and interruption process.

Feature selection [6] is also known as algorithms for evaluating attributes and algorithms for sub-set evaluation. In the first approach features are listed individually and then each feature is assigned a weight based on the degree of importance to the target feature of each feature. In contrast, the second approach selects and ranks feature sub-sets based on certain criteria for evaluation. Attribute assessment methods do not measure correlations among features so that subsets with redundant features are likely to yield. Subset assessment approaches are better at eliminating redundant characteristics. Different types of algorithms for feature selection were suggested. Filter methods, wrapper methods and embedded methods are primarily classified into three groups. Each algorithm for feature selection uses one of three techniques for feature selection. In this article, we propose the optimal feature selection algorithm for the classification of medical datasets based on differential evolution and filter-based selection techniques.

3. RELATED WORKS

De Silva, Kushan, Daniel Jönsson, and Ryan T. Demmer [7]This Work has shown the importance of integrating the collection of features with machinery to recognise a wide variety of predictors that can improve prediabetes and clinical decision making. Training data with 156 pre-selected exposure variables have been used to pick the feature algorithms. In original and re-sampling training data sets with 4 resampling methods, four machine learning algorithms were applied to 46 exposure variables.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 3892 - 3913
Received 20 January 2021; Accepted 08 February 2021.

Christo, VR Elgin, et al [8]This work used the wrapper technique, which uses co-operative co-development and random forest classification, to pick features and instances. The reduced data set is used as a way of training a random classification of forests. These judgments support doctors for diagnosis and care as the second opinion. The University of California Irvine (UCC) Machine Learning library is used for studies in Diagnostic Breast Cancer (WDBC), hepatitis, Pima Indian Diabetes (PID), Cleveland Heart Disease (CHD), Statlog Heart Disease (SHD).

Gandhi, Kriti, et al [9] Machine learning features applied in a specific framework in healthcare facilities. Instead of care for the patient directly, the whole therapy process can be made even more effective if the condition is predicted in advance with some machine learning algorithms. Some cases often happen when a disease is not diagnosed or performed early.

de Lima, Márcio Dias, Juliana de Oliveira Roque e Lima, and Rommel M. Barbosa [10] A new selection algorithm was presented. The authors used 8 benchmark data sets to validate our research, widely used by scientists who developed machine learning methods for the classification of medical data. The experiment demonstrated that the performance of our proposed new selection method, combined with the FSTBSVM, is highly efficient.

Sahebi, Golnaz, et al [11] Proposed a general set of wrappers based on a parallel new Genetic Algorithm (GA) called GeFeS. The proposed GeFeS performs well under various dimensions and sizes of numerical data sets, attempts to prevent overfitting and significantly improves classification precision. A new operator for weighting, improvement in mutation and crossover operators and integrated nested cross validation in the GA process has been suggested in order to make the GA specificand smart to validate the apprenticeship model properly. In order to determine the goodness of selected features, the K-nearest neighbour (kNN) classification is used.

Muthulakshmi, I [12] The CKD classification model for optimum feature selection was presented. The Particle Swam Optimization (PSO) is used for feature selection purposes and medical data classification is carried out by using the Ant-Colony Optimization (ACO) algorithm. The proposed model is tested with a CKD benchmark in multiple stages.

Jain, Divya, and Vijendra Singh [13] In this report, the diagnosis of chronic diseases is provided with a fast and new adaptive classification system. To that end, the proposed solution uses a hybrid approach consisting of the PCA and Relief system with an optimised

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 3892 - 3913
Received 20 January 2021; Accepted 08 February 2021.

support classifier of the vector machine. The SVM classification uses an effective parameter optimization approach to achieve high classification precision, understandability and consistency.

Xie, Jingui, et al. [14] This study focused on the collection and consistency of the definition of syndrome, essential characteristics of demographical detail, personal medical background and symptoms. Methods of collection and classification of mine data on TCM syndromes were employed. The selection of features enhanced classification models efficiency.

Mezzatesta, Sabrina, et al. [15] End-stage kidney disease (ESKD) patients are at unique risk for cardiovascular disease. The purpose of this research was to predict death and cardiovascular diseases in patients with dialysis with a certain accuracy. In particular, the authors have achieved optimum efficiency with Grid Search using the non-linear SVC with the RBF kernel algorithm. The above is an algorithm that helps to look for the best hyper parameters combination (in which case to find the best pair to increase the algorithm 's precision).

Álvarez, Josefa Díaz, et al [16]Five algorithms were applied to evaluate the related 18F-fluorodeoxyglucose positron tomography characteristics of major areas of PPA affected in patient records. On the other hand, before and after the selection process, we conducted classification and clustering algorithms in comparison to the results achieved in previous works. In order to further give a system for automated diagnostic aid, we tried to find the best classifier and more appropriate features in the WEKA method.

Toğaçar, M., et al [17] Utilized images of lung X-ray for pneumonia diagnosis. The convolutional neural network was used as a feature extractor to perform this special function by using some of the current convolutional neural network models AlexNet, VGG-16 and VGG-19. Then, the number of deep characteristically, the overall redundancy algorithm for each deep model has been reduced from 1000 to 100.

Cömert, Zafer, et al [18] In this article two wrapper, and three philtres feature selection methods and machine learning models are tested on high-dimensional feature sets obtained from open access CTU-UHB intrapartum CTG databases. The models are the artificial neural network (ANN) k-nearest neighbour (k-NN), decision Tree (DT) and Supporting Vector Machine (SVM).

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 3892 - 3913
Received 20 January 2021; Accepted 08 February 2021.

Raihan-Al-Masud, Md, and M. Rubaiyat Hossain Mondal [19] Concentrated on the use of spinal abnormality prediction algorithms for machine learning. As a data pre-processing step, a single feature selection is considered as a philtre selection, and the main component analysis (PCA) is considered as a feature extraction algorithm. A variety of machine-learning methods are considered for diagnosis of spinal abnormality such as SupportVector Machine (SVM), Logistic Regression (LR).

Özyurt, Fatih [20] White blood cell (WBC) Testing is used to diagnose many diseases, including leukaemia allergies, particularly infections. To identify and recognise the quantity of WBCs in human blood, a doctor requires clinical experiences. Eosinophils, lymphocytes, monocytes and neutrophils are classified into four subclasses. The research has included pre-trained architectures such as AlexNet, VGG-16, Google Net and ResNet. The characteristics obtained in these architectures' last completely connected layers were combined. The minimum redundancy maximum pertinence approach used was used to pick productive characteristics. In comparison to classical neural network (CNN) architectures, thanks to the efficient features obtained from CNN architectures the extreme learning engine (ELM) classification was implemented in the classification process.

Li, Jian Ping, et al [21] The technology is based on machine learning techniques and is proposed an accurate and exact diagnostic system for heart disease. It is based on classification algorithms that include the vector system, the logistic regression, the artificial neural network, the nearest K neighbour, Naïve Bays and Decision Tree. The standard features include relief, maximum redundancy, the least absolute retreat, and local learning for redundancy removal the model was developed based on classification algorithms.

4. PROPOSED OPTIMAL FEATURE SELECTION TECHNIQUE FOR MEDICAL DATASETS

In this proposed technique, the detailed description of the filter-based feature selection techniques and DF optimization algorithm are given. In this proposed approach, the best and worst solutions are obtained by converting the real code into binary values string to speed up the process, for reducing the computation time.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 3892 - 3913
Received 20 January 2021; Accepted 08 February 2021.

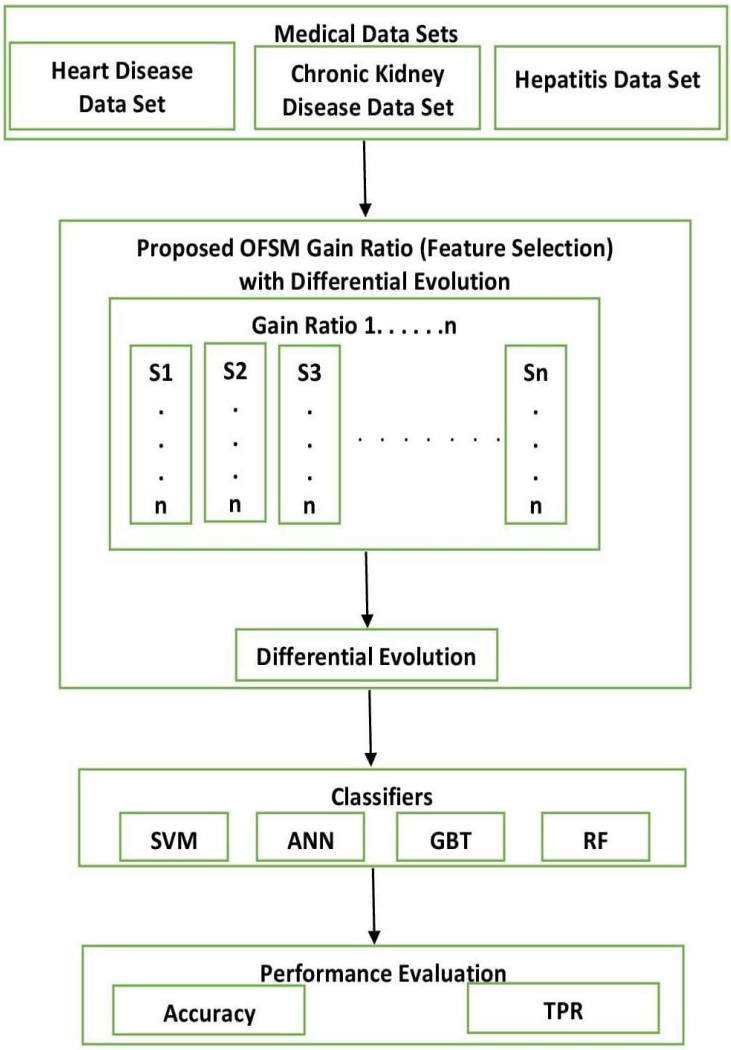

Figure 1: Proposed Methodology

4.1 Gain Ratio Feature Selection

The Gain Ratio [22] is the non-symmetrical measure that is presented to pay back on the bias of the Information Gain (IG). GR is given by Equation (1):

$$GR = \frac{Information\ Gain(IG)}{H(X)} \quad (1)$$

Information Gain (IG) is a symmetrical measure.

$$IG = H(Y) - H(Y|X) = H(X) - H(X|Y) \quad (2)$$

The information gained about Y after observingX is alike to the information gained about X afterobserving Y in the Equation (2). There, a weaknessof the IG criterion is that it is biased in favor offeatures with more values even when they are notmore informative.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 3892 - 3913
Received 20 January 2021; Accepted 08 February 2021.

As in the above Equation (2) presents, when the variable Y has to be predicted, then regularize the IG by distributing the entropy of X, and vice versa. Owing to this normalization, the GR values constantly fall in the range [0, 1]. A value of GR = 1 specifies that the knowledge of X totally forecasts Y, and GR = 0 means that there is no relation between Y and X. In opposition to IG, the GR support variables with lesser values.

4.3 Differential Evolution Optimization

Differential evolution (DE) is merely one of several approaches through evolutionary algorithm where in actuality the features are search and centred on ant colony. An easy and yet effective, DE give you the benefits usually requires like many optimization methods[23][24]. There are several actions from DE such; 1) ability to handle non-differentiable, nonlinear and value this is certainly multimodal, 2) parallelizability to cope with computation cost that is intensive, 3) simplicity of good use, 4) good convergence properties.

Like GA, DE employ factors which can be same of mutation, selection and crossover. The efficiency of DE depends on the handling of target vector and difference in order to acquire a task vector in exploring procedure. Every real-value this is certainly d-Dimensional, a population of NP members is provided. NP will be the population size and D will be the true range that is wide of to be fine-tuned. Among the members of two population like y_{s2} and y_{s3} added the vector of weight difference to the y_{s1} which is third member for creating a trial vector. This action is termed as mutation. A mutant vector is generating relating to for every target vector$y_{(I,G)}, j = 1,2,3, \dots, M$a mutant vector using the given equation:

$$w_{j,G+1} = y_{s1,H} + G\left(y_{s2,H} - y_{s2,H}\right) \qquad (2)$$

Where $s_1, s_2, s_3 \in \{1,2, \dots, NP\}$are integers that are chosen randomly, should be specific from 1 another plus unique through the operating index j. The control rate of Scaling factor F(0,1) that your particular population comprises. In order to improve the variety in connection with perturbed factor vectors, introduction of crossover is takes place. The trial vector:

$$v_{j,H+1} = \left(v_{1,j,H+1}, v_{2,j,H+1}, \dots, v_{E,j,H+!}\right) \qquad (3)$$

Is from where;

$$v_{kj,H+1} = \begin{cases} w_{kj,H+1} \; if \; rand(0,1) \leq d_s \\ y_{kj,H+1} \; otherwise \end{cases} \qquad (4)$$

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 3892 - 3913
Received 20 January 2021; Accepted 08 February 2021.

Where the H is the current population and the trial vector $k^{th}j^{th}$for the dimension of $v_{(kj,H)}$. The probability of crossover$d_s(0,1)$ is a person described value that operates the portion in connection with parameter values which are often and that can be replicated through the mutant. Selection will be the stage to get the vector among the target vector as well as trial vector making use of the aim of generating an individual in terms of generation this is certainly next. Then your causing vector substitutes the vector with which it absolutely was compared [25] if the recently created vector leads to a lower objective feature value (better fitness) as compared to population member that is predetermined. But, many factors from DE are instantly transformative without needed user to see by learning from your own error's strategy. In this work that is ongoing size of generation and population are adaptively identifying predicated on a total of features remained from relief-f. Hence, the buyer doesn't always have to initialize those factor values manually.

4.3 Proposed Optimal Feature Selection Method

In this proposed OFS method, instead of using Crossover and Mutation operator of DF optimization, encoding of solution (converting real code to binary string) is introduced to consume the computation time for medical datasets during the classification of diseases. The following are stages involved in this proposed OFS method is given below:

Stage 1: Encoding of Solution

Each individual solution is expressed in this work as a binary string in the population. The length of each solution (binary string) is equal to that of different features in the medical data sets. The solution's binary code 1 indicates the feature selection and the solution's binary code 0 is the feature not selected. The $S = [F_1, F_2, F_3, .. F_m]$ is the solution where m is the different datasetfeatures. For example, a solution described as a [1, 1, 0 , 0, 1 , 1, 1, 0 , 0 , 0, 0 , 1] feature with index 1, 2 , 5, 6 and 10 is selected while the other features are not selected. Each location in the solution is binary value, q as a solution. The generationi of p^{th} solution is represented with $S_p^{(i)}$ and $S_{p,q}^{(i)}$ is represented by q^{th}position.

Stage 2: Initial population

Set the population size for this project to be 50. We produce 50 random solutions with real values randomly varying from 0 to 1. After that, each individual solution is used to convert real values to binary values based on the following equation: The digitization process:

$$S_{p,q}^{(i)} = \begin{cases} 1 & S_{p,q}^{(i)} > rand \\ 0 & Otherwise \end{cases}$$

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 3892 - 3913
Received 20 January 2021; Accepted 08 February 2021.

Rand is a random number distributed uniformly from 0 to 1.

Stage 3: Fitness function

In optimization problems, fitness plays a key role. A fitness function measures the output of a single positive integer. With the aid of classifier error rate formulated in the following equations, the fitness for each solution in the population is calculated:

$$fitness\left(S_p^{(i)}\right) = ClassifierErrorRate\left(S_p^{(i)}\right)$$

The classifier error rate (solution) is the testing error rate:

$$ClassifierErrorRate\left(S_p^{(i)}\right) = \frac{Number\ of\ misclassified\ records}{Total\ number\ of\ records} \text{X } 100$$

Stage 4: Finding new solutions

To produce the new solution the best and worst solutions in the output t are used. The lowest fitness value of the best solution (error rate) and the lowest fitness value of the generation i. $S_{bt}^{(i)}$ is represents an i iteration best solutionand the $S_{wt}^{(i)}$is used to represent the worst solution. Taking into account the best and worst generation solutions 'i', the q^{th} position of old solution $S_{p,q}^{(i)}$ as formulated by the below equation:

$$S_{p,q}^{(i)} = S_{p,q}^{(i)} + A\left|S_{bt,q}^{(i)} - S_{p,q}^{(i)}\right| + B\left|S_{wt,q}^{(i)} - S_{p,q}^{(i)}\right|$$

In case A, B is between 0 and 1 random numbers. Afterwards, the digitalization process is used in the next generation 'i+1' to transform real values into binary values based on the following equation for each position of the candidate:

$$S_{p,q}^{(i)} = \begin{cases} 1, & S_{p,q}^{(i+1)} > rand \\ 0, & Otherwise \end{cases}$$

Here rand is a uniformly distributed random numberbetween 0 and 1.

Stage 5: Termination criteria

An iterative process is the proposed work. By the following considerations the termination conditions for the iterative process can be determined:

- Total number of iterations (T_{max})
- Fitness convergence rate and
- Threshold for iterative process running time In this proposed work, the maximum number of iterations (T_{max}) is used as the termination criterion.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 3892 - 3913
Received 20 January 2021; Accepted 08 February 2021.

Step by Step procedure for Proposed Optimal Feature Selection Method

Input: Medical Datasets (MD)

Output: Optimal Feature Subset (selected best features)

Step 1: Splitting of the Dataset into Training and Test (MD = $MD_{tr} + MD_{ts}$)

Step 2: Applying T = GR ←MD

Step 3: $m_f = |T|$

Step 4: Constructing Initial Population Table

Step 4.1: foreach solution S_p p= 1 to M do

Step 4.1.1: foreach position q of solution S_p; $q = 1$ to m_f do

Step 4.1.2: $S_{p,q} = rand(0,1)$

Step 4.1.3: $S_{p,q} = Digitization(S_{p,q})$

Step 4.1.4: end

Step 4.2: $S_p^{fitness} = computeFitness(S_p, C, f, MD_{tr}, MD_{ts})$

Step 4.3: end

Step 5: $S_{bt} = findBestSolution()$

Step 6: $S_{wt} = findWorstSolution()$

Step 7: Iterative Process

Step 7.1: foreach iteration i 1 to T_{max} do

Step 7.1.1: foreach position q of solution S_p $p = 1\ to\ M$ do

Step 7.1.1.1: foreach position q of solution S_p; $q = 1\ to\ m_f$ do

Step 7.1.1.2: $A = rand(0,1);\ B = rand(0,1)$

Step 7.1.1.3: $F_{p,q} = S_{p,q}^{(i)} + A\left|S_{bt,q}^{(i)} - S_{p,q}^{(i)}\right| + B\left|S_{wt,q}^{(i)} - S_{p,q}^{(i)}\right|$

Step 7.1.1.4: $F_{p,q} = Digitization(F_{p,q})$

Step 7.1.1.5: end

Step 7.1.2: $F_p^{fitness} = computeFitness(F_p, C, f, MD_{tr}, MD_{ts})$

Step 7.1.3: if $F_p^{fitness} < S_p^{fitness}$ then

Step 7.1.3.1: $S_p = F_p$

Step 7.1.4: end

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 3892 - 3913
Received 20 January 2021; Accepted 08 February 2021.

Step 7.2: end

Step 7.3: $S_{bt} = findBestSolution()$

Step 7.4: $S_{wt} = findWorstSolution()$

Step 8: end

Step 9: Extracting the best optimal feature subset from the Ψ_{best}

Step 10: foreach position j of solution S_{bt} j=1 to m_f do

Step 10.1: if isPositionSelected($S_{bt,q}$) then

Step 10.1.1: $OF_{bt} = OF_{bt} \cup T[q]$

Step 10.2: end

Step 11: end

Return OF_{bt}

5. RESULT AND DISCUSSION

The medical datasets like Heart Disease [26], Chronic Kidney Disease [27], and Hepatitis [28] are considered from the public UCI and Kaggle Repositories. The performance metrics like Accuracy (in %), and True Positive Rate (in %)are considered in this research work.

5.1 Number of Features obtained

Table 1 depicts the number of features obtained by Gain Ratio, Information Gain, Differential Evolution, Genetic Algorithm. From the table 1, the proposed Optimal Feature Selection method generates least number of features than other existing feature selection techniques.

Table 1: Number of Features obtained by Feature Selection Techniques for Medical datasets like Heart Disease, Kidney Disease and Hepatitis

Feature Selection Techniques	Number of Features obtained		
	Heart Disease	Kidney Disease	Hepatitis
Original dataset	13	24	19
Gain Ratio	9	13	15
Information Gain	10	22	17
Differential Evolution	7	15	13

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 3892 - 3913
Received 20 January 2021; Accepted 08 February 2021.

Genetic Algorithm	12	21	16
Proposed Optimal Feature Selection Method	5	11	9

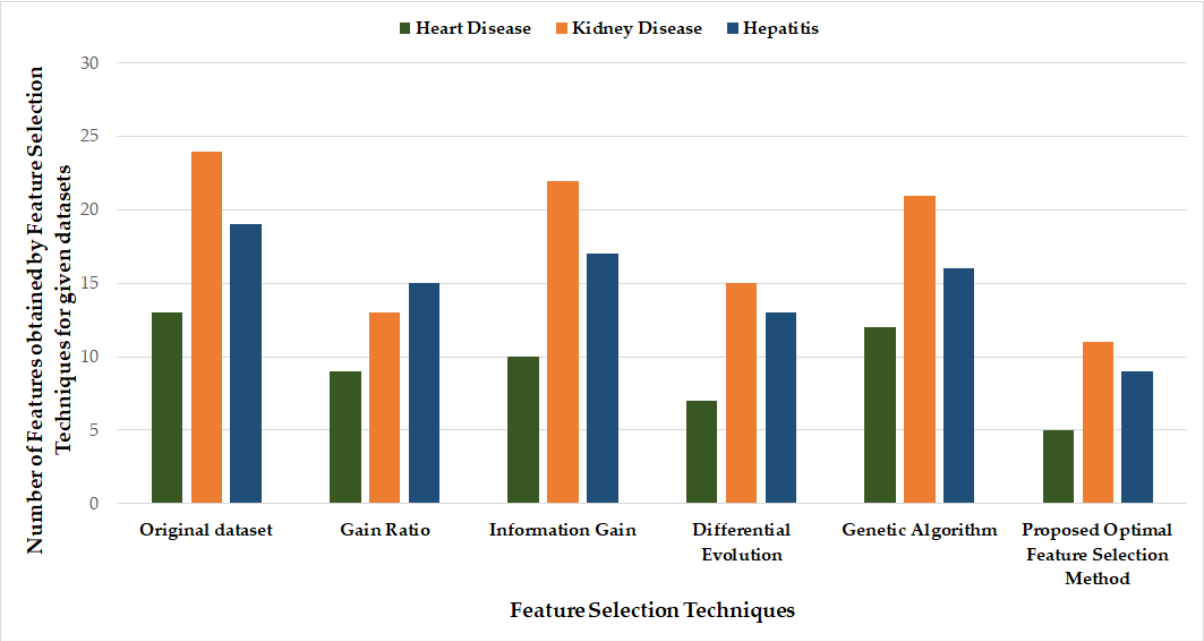

Figure 1: Graphical representation of the number of features obtained by Existing and Proposed Feature Selection techniques for given three datasets

5.3 Classification Accuracy (in %)

The performance analysis of the feature selection techniques like GR, IG, DE, GA, and Proposed OFSM are analysis with the classifiers like Random Forest, Gradient Boosting Tree, Artificial Neural Network and Support Vector Machine. Table 2.1 gives the Accuracy (in %) for the Heart disease (HD)dataset obtained by the various classifiers like RF, GBT, ANN, and SVM with Original dataset, GR, IG, DE, GA and Proposed OFSM. Figure 2.1 depicts the graphical representation of the Accuracy (in %) obtained for the Heart Disease dataset with various classifiers using GR, IG, DE, GA and Proposed OFS method. From the table 2.1 and figure 2.1, the original HDdataset, GR processed HD dataset, DE processed HD dataset and proposed OFSM processed HD dataset with SVM classifiers gives increased accuracy than the other feature selection techniques with other classifiers.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 3892 - 3913
Received 20 January 2021; Accepted 08 February 2021.

Table 2.1: Accuracy (in %) obtained for the Heart Disease dataset with various classifiers using GR, IG, DE, GA and Proposed OFS method

Feature Selection Methods	Accuracy (in %) Classification Techniques			
	RF	GBT	ANN	SVM
Original Dataset	43.75	48.16	46.51	**49.175**
GR processed dataset	66.67	67.76	69.73	**69.92**
IG processed dataset	61.16	64.64	**66.57**	64.32
GA processed dataset	61.41	64.59	**67.71**	65.31
DE processed dataset	68.92	71.19	73.69	**74.68**
Proposed OFSM processed dataset	77.71	92.76	93.93	**95.47**

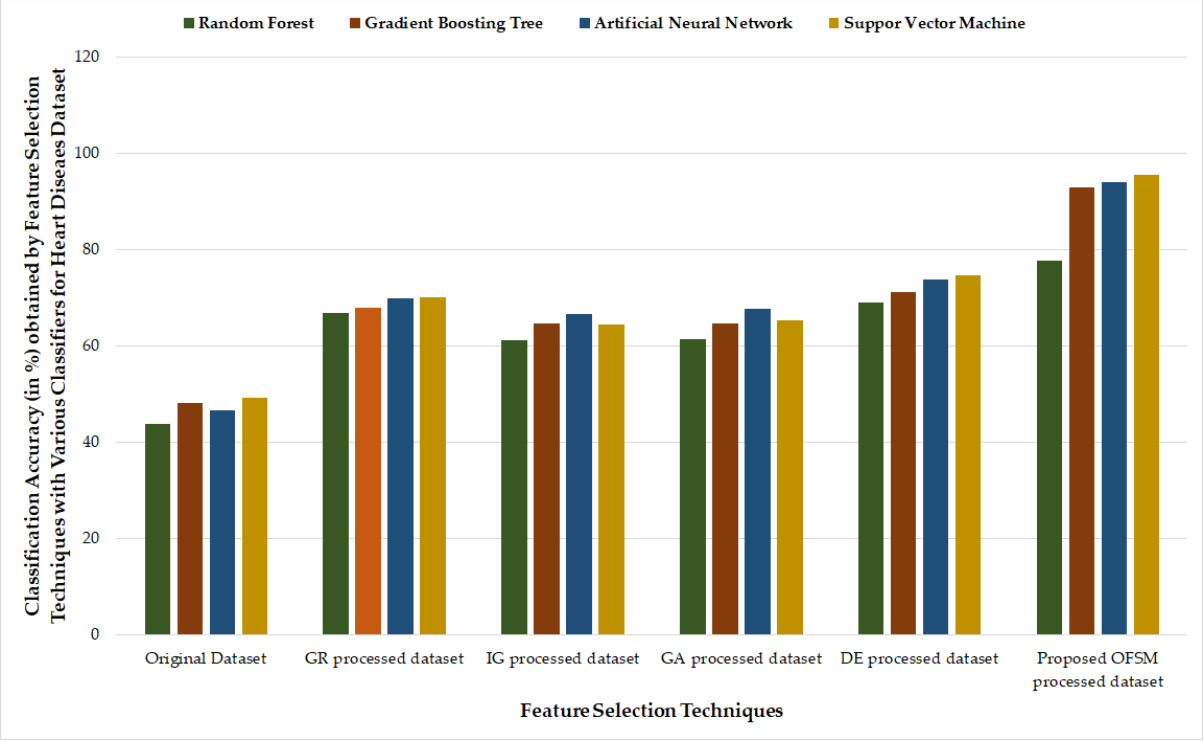

Figure 2.1: Graphical representation of the Classification Accuracy (in %) obtained by feature selection techniques with various classifiers for Heart Disease dataset

Table 2.2 gives the Accuracy (in %) for the Chronic Kidney disease dataset obtained by the various classifiers like RF, GBT, ANN, and SVM with Original dataset, GR, IG, DE, GA and Proposed OFSM. Figure 2.2 depicts the graphical representation of the Accuracy (in %) obtained for the Chronic Kidney Disease (CKD) dataset with various classifiers using GR, IG, DE, GA and Proposed OFS method. From the table 2.2 and figure 2.2, it is clear that the

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 3892 - 3913
Received 20 January 2021; Accepted 08 February 2021.

Original CKD dataset, GR, IG, DE, and GA processed CKD dataset with SVM classifiers gives improved accuracy than the other classifiers.

Table 2.2: Accuracy (in %) obtained for the Chronic Kidney Disease (CKD) dataset with various classifiers using GR, IG, DE, GA and Proposed OFS method

Feature Selection Methods	Accuracy (in %) Classification Techniques			
	RF	GBT	ANN	SVM
Original Dataset	44.75	46.76	42.86	**47.13**
GR processed dataset	68.53	69.72	67.43	**72.65**
IG processed dataset	59.61	60.67	60.54	**61.42**
GA processed dataset	58.52	59.96	59.86	**60.31**
DE processed dataset	72.79	72.65	71.37	**79.96**
Proposed OFSM processed dataset	91.64	93.27	93.72	**98.58**

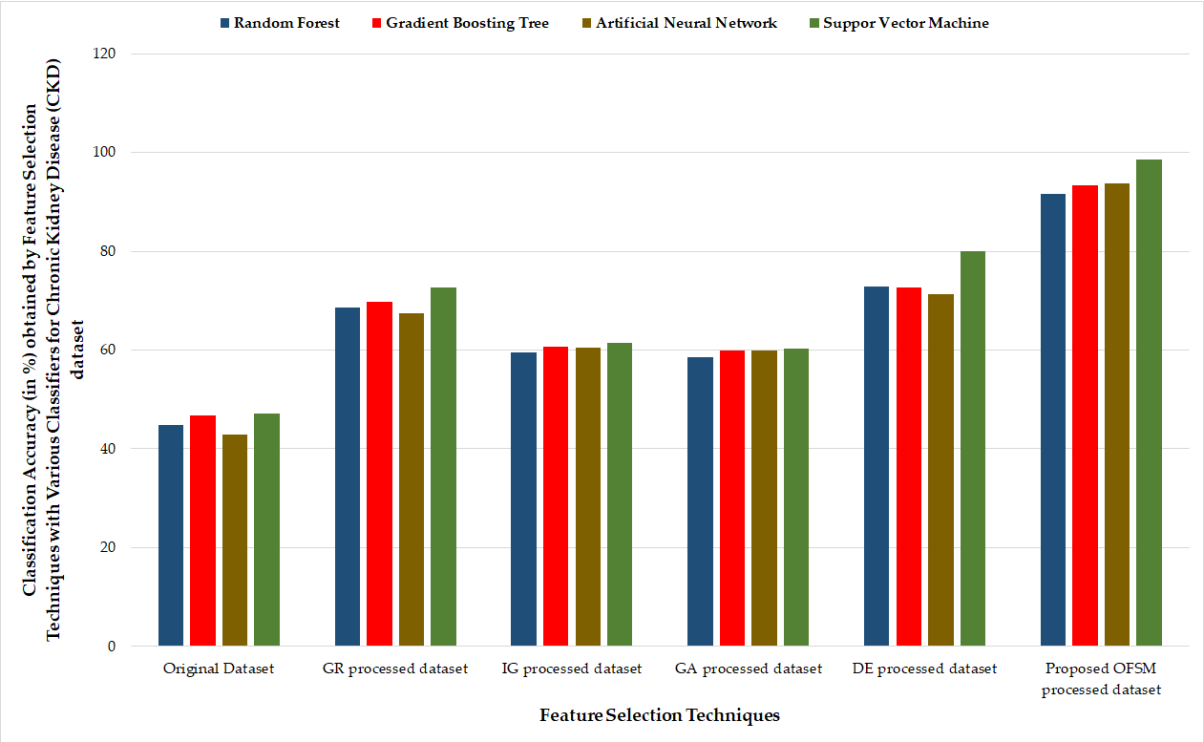

Figure 2.2: Graphical representation of the Accuracy (in %) obtained for the Chronic Kidney Disease (CKD) dataset with various classifiers using GR, IG, DE, GA and Proposed OFS method

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 3892 - 3913
Received 20 January 2021; Accepted 08 February 2021.

Table 2.3 gives the Accuracy (in %) for the Hepatitis disease dataset obtained by the various classifiers like RF, GBT, ANN, and SVM with Original dataset, GR, IG, DE, GA and Proposed OFSM.Figure 2.3 depicts the graphical representation of the Accuracy (in %) obtained for the Hepatitis (HP) Disease dataset with various classifiers using GR, IG, DE, GA and Proposed OFS method. From the table 2.3 and figure 2.3, it is clear that the Original HP dataset, GR processed HP, DE Processed HP dataset and Proposed OFSM processed HP dataset with ANN gives improved accuracy, and IG processed HP, and GA processed HP dataset gives increased accuracy with RF classifier when it is compared with other classifiers.

Table 2.3: Accuracy (in %) obtained for the Hepatitis (HP) Disease dataset with various classifiers using GR, IG, DE, GA and Proposed OFS method

Feature Selection Methods	Accuracy (in %) Classification Techniques			
	RF	GBT	ANN	SVM
Original Dataset	45.75	46.63	**51.34**	44.64
GR processed dataset	69.63	69.34	**70.27**	65.46
IG processed dataset	**59.74**	59.46	56.46	54.59
GA processed dataset	**57.63**	56.51	55.37	53.28
DE processed dataset	73.16	71.42	**74.17**	70.27
Proposed OFSM processed dataset	93.57	92.63	**95.15**	90.43

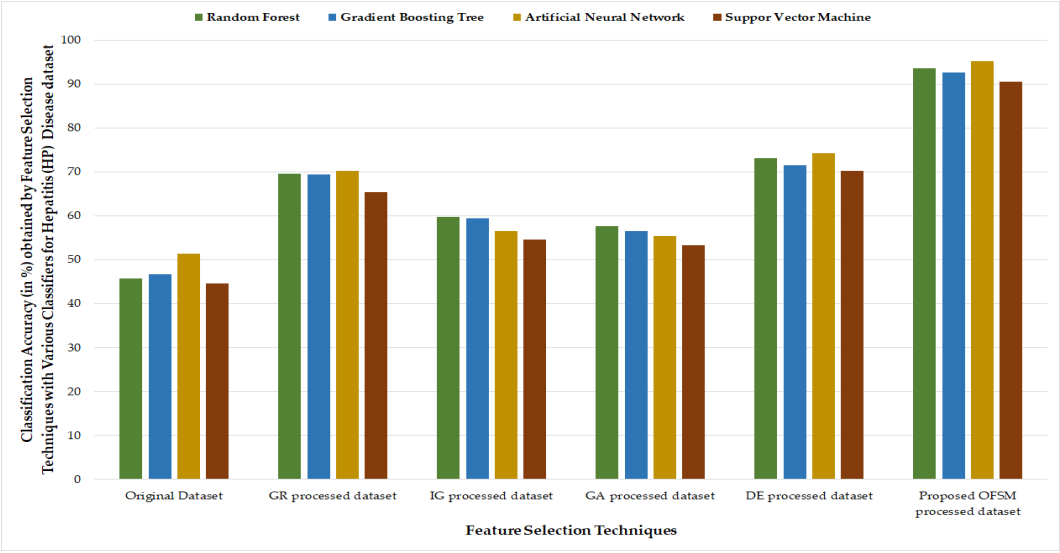

Figure 2.3: Graphical representation of the Accuracy (in %) obtained for the Hepatitis (HP) Disease dataset with various classifiers using GR, IG, DE, GA and Proposed OFS method

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 3892 - 3913
Received 20 January 2021; Accepted 08 February 2021.

5.4 True Positive Rate (in %)

Table 3.1 gives the True Positive Rate (in %) for the Heart disease (HD) dataset obtained by the various classifiers like RF, GBT, ANN, and SVM with Original dataset, GR, IG, DE, GA and Proposed OFSM.Figure 3.1 depicts the graphical representation of the True Positive Rate (in %) obtained for the Heart Disease dataset with various classifiers using GR, IG, DE, GA and Proposed OFS method. From the table 3.1 and figure 3.1, it is clear that the original HD dataset, GR, IG, GA, DE, and Proposed OFSM processed HD dataset with SVM classifiers gives more TPR rate than the other classifiers.

Table 3.1: True Positive Rate (in %) obtained for the Heart Disease dataset with various classifiers using GR, IG, DE, GA and Proposed OFS method

Feature Selection Methods	TPR (in %) by Classification Techniques			
	RF	GBT	ANN	SVM
Original Dataset	53.43	53.76	53.62	**55.50**
GR processed dataset	74.15	72.77	70.17	**75.41**
IG processed dataset	68.06	66.46	66.23	**69.54**
GA processed dataset	69.56	69.79	69.37	**72.42**
DE processed dataset	74.55	73.59	71.76	**75.54**
Proposed OFSM processed dataset	91.57	93.72	93.85	**95.84**

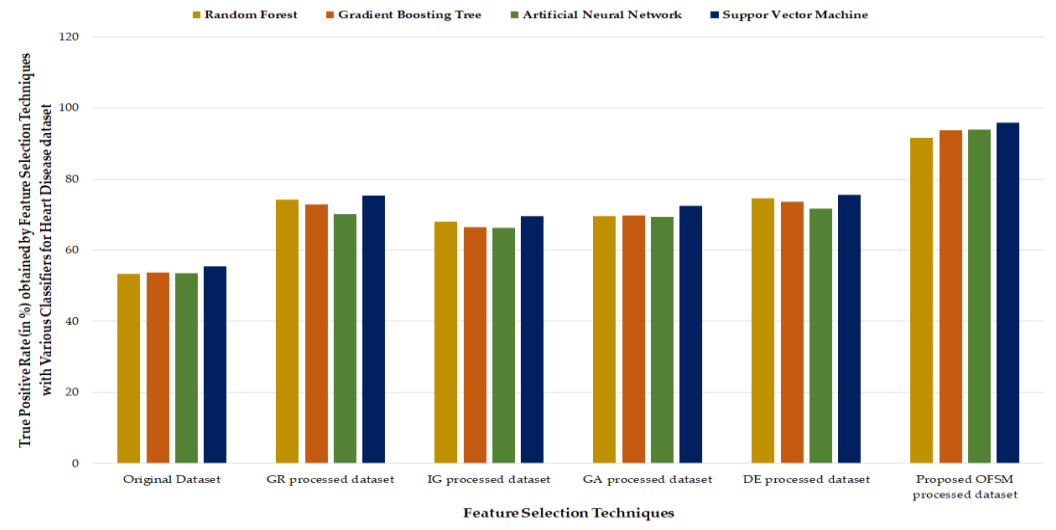

Figure 3.1: Graphical representation of the True Positive Rate (in %) obtained for the Heart Disease dataset with various classifiers using GR, IG, DE, GA and Proposed OFS method

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 3892 - 3913
Received 20 January 2021; Accepted 08 February 2021.

Table 3.2 gives the True Positive Rate (in %) for the Chronic Kidney Disease (CKD) dataset obtained by the various classifiers like RF, GBT, ANN, and SVM with Original dataset, GR, IG, DE, GA and Proposed OFSM. Figure 3.2 depicts the graphical representation of the True Positive Rate (in %) obtained for the Chronic Kidney Disease dataset with various classifiers using GR, IG, DE, GA and Proposed OFS method. From the table 3.2 and figure 3.2, it is clear that the original CKD dataset, GR, DE, and Proposed OFSM processed CKD dataset with SVM classifiers gives more TPR rate than the other classifiers.

Table 3.2: True Positive Rate (in %) obtained for the Chronic Kidney Disease dataset with various classifiers using GR, IG, DE, GA and Proposed OFS method

Feature Selection Methods	TPR (in %) by Classification Techniques			
	RF	GBT	ANN	SVM
Original Dataset	52.44	48.86	47.56	**53.94**
GR processed dataset	71.83	74.31	73.27	**74.98**
IG processed dataset	64.37	65.63	**64.15**	62.36
GA processed dataset	62.45	63.43	**64.54**	60.35
DE processed dataset	70.46	73.72	72.37	**81.24**
Proposed OFSM processed dataset	91.61	93.73	93.32	**95.78**

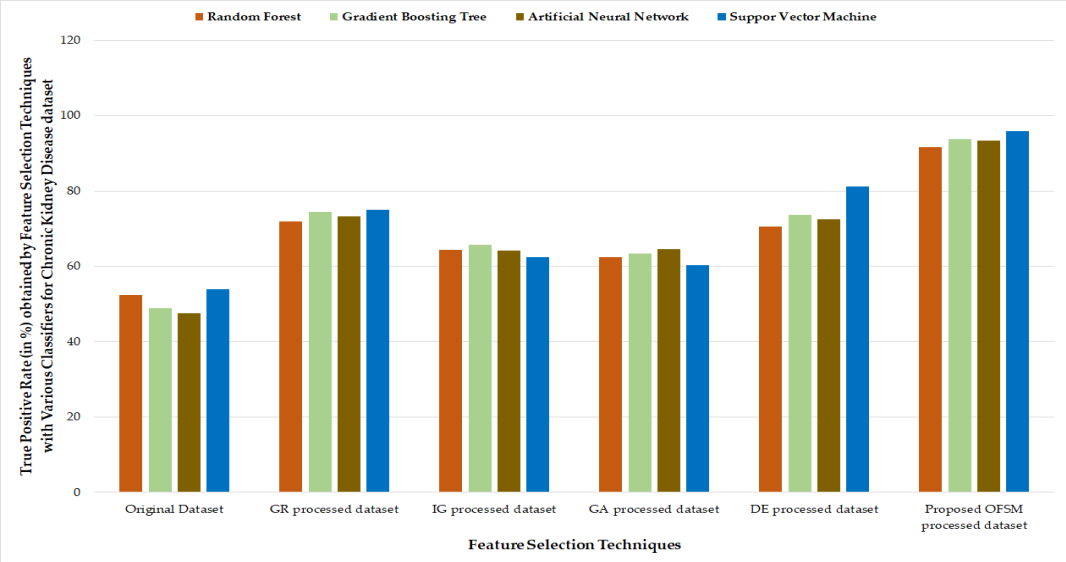

Figure 3.2: Graphical representation of the True Positive Rate (in %) obtained for the Chronic Kidney Disease dataset with various classifiers using GR, IG, DE, GA and Proposed OFS method

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 3892 - 3913
Received 20 January 2021; Accepted 08 February 2021.

Table 3.3 gives the True Positive Rate (in %) for the HepatitisDisease (HP) dataset obtained by the various classifiers like RF, GBT, ANN, and SVM with Original dataset, GR, IG, DE, GA and Proposed OFSM. Figure 3.3 gives the graphical representation of the True Positive Rate (in %) obtained for the Hepatitis Disease (HP) dataset with various classifiers using GR, IG, DE, GA and Proposed OFS method. From the table 3.3 and figure 3.3, it is clear that the original HP dataset, GR, DE, GA and Proposed OFSM processed HP dataset with ANN classifiers gives more TPR rate than the other classifiers.

Table 3.3: True Positive Rate (in %) obtained for the Hepatitis Disease (HP) dataset with various classifiers using GR, IG, DE, GA and Proposed OFS method

Feature Selection Methods	TPR (in %) by Classification Techniques			
	RF	GBT	ANN	SVM
Original Dataset	48.71	50.62	**54.73**	53.18
GR processed dataset	65.47	67.31	**69.78**	63.28
IG processed dataset	55.31	57.52	55.62	**61.64**
GA processed dataset	56.43	58.71	**60.47**	59.92
DE processed dataset	82.92	81.75	**85.73**	79.84
Proposed OFSM processed dataset	94.31	94.14	**95.72**	90.77

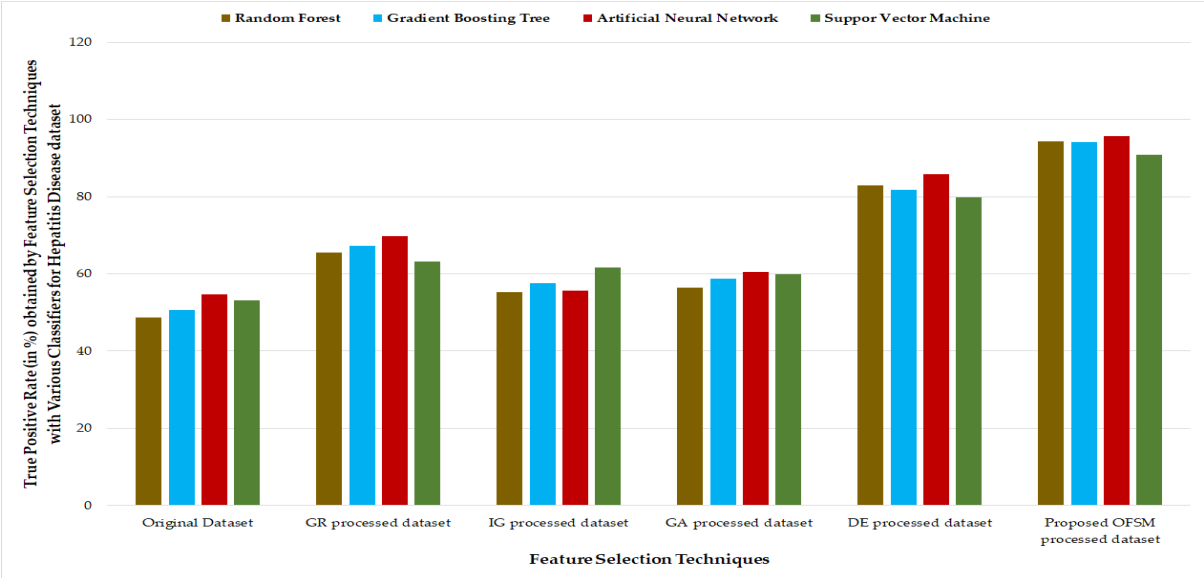

Figure 3.3: Graphical representation of the True Positive Rate (in %) obtained for the Hepatitis Disease (HP) dataset with various classifiers using GR, IG, DE, GA and Proposed OFS method

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 3892 - 3913
Received 20 January 2021; Accepted 08 February 2021.

6. CONCLUSION

In medical diagnosis, it is very important to identify most significant risk factors related to disease. Relevant feature identification helps in the removal of unnecessary, redundant attributes from the disease dataset which, in turn, gives quick and better results. In this research work, an optimization-based feature selection with Filter approach is proposed to enhance the classification accuracy of the disease diagnosis. The combination of GR filter-based feature selection with DE optimization to find the most relevant features for the disease detection for various medical datasets like Heart Disease, Chronic Kidney Disease, and Hepatitis disease. From the result obtained, it is clear that the proposed OFSM with SVM classifiers performs better for HD dataset and CKD dataset, whereas proposed OFSM with ANN gives better result for Hepatitis dataset.

REFERENCES

[1] Weiss, Sholom M., et al. "A model-based method for computer-aided medical decision- making." *Artificial intelligence* 11.1-2 (1978): 145-172.

[2] Iswanto, Iswanto, et al. "Identifying diseases and diagnosis using machine learning." (2019).

[3] Rani, Rella Usha, and Jagadeesh Kakarla. "Efficient Classification Technique on Healthcare Data." *Progress in Advanced Computing and Intelligent Engineering*. Springer, Singapore, 2019. 293-300.

[4] Nishant, Potnuru Sai, et al. "Identifying Classification Technique for Medical Diagnosis." *ICT Analysis and Applications*. Springer, Singapore, 2020. 95-104.

[5] Al-Tashi, Qasem, et al. "A review of grey wolf optimizer-based feature selection methods for classification." *Evolutionary Machine Learning Techniques*. Springer, Singapore, 2020. 273-286.

[6] Sánchez-Reyna, Ana Gabriela, et al. "Feature Selection and Machine Learning Applied for Alzheimer's Disease Classification." *Latin American Conference on Biomedical Engineering*. Springer, Cham, 2019.

[7] De Silva, Kushan, Daniel Jönsson, and Ryan T. Demmer. "A combined strategy of feature selection and machine learning to identify predictors of prediabetes." *Journal of the American Medical Informatics Association* 27.3 (2020): 396-406.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 3892 - 3913
Received 20 January 2021; Accepted 08 February 2021.

[8] Christo, VR Elgin, et al. "Feature Selection and Instance Selection from Clinical Datasets Using Co-operative Co-evolution and Classification Using Random Forest." *IETE Journal of Research* (2020): 1-14.

[9] Gandhi, Kriti, et al. "Disease Prediction using Machine Learning."

[10] de Lima, Márcio Dias, Juliana de Oliveira Roque e Lima, and Rommel M. Barbosa. "Medical data set classification using a new feature selection algorithm combined with twin-bounded support vector machine." *Medical & Biological Engineering & Computing* 58.3 (2020): 519-528.

[11] Sahebi, Golnaz, et al. "GeFeS: A generalized wrapper feature selection approach for optimizing classification performance." *Computers in biology and medicine* 125 (2020): 103974.

[12] Muthulakshmi, I. "OPTIMAL FEATURE SELECTION BASED DATA CLASSIFICATION MODEL FOR CHRONIC KIDNEY DISEASE PREDICTION."

[13] Jain, Divya, and Vijendra Singh. "A two-phase hybrid approach using feature selection and adaptive SVM for chronic disease classification." *International Journal of Computers and Applications* (2019): 1-13.

[14] Xie, Jingui, et al. "Feature selection and syndrome classification for rheumatoid arthritis patients with Traditional Chinese Medicine treatment." *European Journal of Integrative Medicine* 34 (2020): 101059.

[15] Mezzatesta, Sabrina, et al. "A machine learning-based approach for predicting the outbreak of cardiovascular diseases in patients on dialysis." *Computer methods and programs in biomedicine* 177 (2019): 9-15.

[16] Álvarez, Josefa Díaz, et al. "An application of machine learning with feature selection to improve diagnosis and classification of neurodegenerative disorders." *BMC bioinformatics* 20.1 (2019): 491.

[17] Toğaçar, M., et al. "A deep feature learning model for pneumonia detection applying a combination of mRMR feature selection and machine learning models." *IRBM* 41.4 (2020): 212-222.

[18] Cömert, Zafer, et al. "Prediction of intrapartum fetal hypoxia considering feature selection algorithms and machine learning models." *Health information science and systems* 7.1 (2019): 17.

[19] Raihan-Al-Masud, Md, and M. Rubaiyat Hossain Mondal. "Data-driven diagnosis of spinal abnormalities using feature selection and machine learning algorithms." *Plos one* 15.2 (2020): e0228422.

[20] Özyurt, Fatih. "A fused CNN model for WBC detection with MRMR feature selection and extreme learning machine." *Soft Computing* 24.11 (2020): 8163-8172.

[21] Li, Jian Ping, et al. "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare." *IEEE Access* 8 (2020): 107562-107582.

[22] Priyadarsini, R. Praveena, M. L. Valarmathi, and S. Sivakumari. "Gain ratio based feature selection method for privacy preservation." *ICTACT J Soft Comput* 1.04 (2011): 2229-6956.

[23] Sudeeptha, J., and C. Nalini. "Hybrid Optimization of Cuckoo Search and Differential Evolution Algorithm for Privacy-Preserving Data Mining." *International Conference on Artificial Intelligence, Smart Grid and Smart City Applications*. Springer, Cham, 2019.

[24] Liang, Jing, et al. "Multimodal multiobjective optimization with differential evolution." *Swarm and evolutionary computation* 44 (2019): 1028-1059.

[25] Brezocnik, Lucija, IztokFister, and GregaVrbancic. "Applying Differential Evolution with Threshold Mechanism for Feature Selection on a Phishing Websites Classification." *European Conference on Advances in Databases and Information Systems*. Springer, Cham, 2019.

[26] https://archive.ics.uci.edu/ml/datasets/heart+disease

[27] https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease

[28] https://archive.ics.uci.edu/ml/datasets/hepatitis