# Improving the previous results for the prediction of stenosis in coronary arteries using Random Forest with Synthetic Minority Oversampling Technique (SMOTE) filter and Metaphor Search

**Akansha Singh***
**Ashish Payal****

Vivekananda Journal of Research
July - December 2021, Vol. 11, Issue 2, 181-203
ISSN 2319-8702(Print)
ISSN 2456-7574(Online)
Peer Reviewed Refereed Journal
© Vivekananda Institute of Professional Studies
http://www.vips.edu/vjr.php

**ABSTRACT**

*An early prediction of a noxious disease like coronary artery disease (CAD) is crucial since it has caused millions of deaths a year. Even after performing bountiful researches, we are still not able to procure a final feature set for the early prediction of CAD. This study has enlightened two main factors that have much helped in achieving excellent results. They are dataset imbalance problem and metaphor (swarm) searches for feature selection. The dataset imbalance problem is resolved using the Synthetic Minority Oversampling Technique (SMOTE). Also, along with swarm searches, several other feature selection methods have been compared concerning some supervised classifiers. After comparing all the methods, the study shows thatswarm searches and the random forest classifiers have surpassed the other methods. The accuracies obtained for left anterior descending (LAD), left circumflex (LCX), and right coronary artery (RCA) are 97.2%, 95.97%, and 94.48%, respectively.*

**Keywords:** *CAD, LAD, LCX, RCA, SMOTE, CFS, Metaphor Search, Random Forest, Naïve Bayes.*

## INTRODUCTION

CAD is one of the deadliest cardiovascular diseases (CVD), causing death all over the world. The blockage of oneor more of the three major arteries pumping blood to the

* Department of Computer Science and Engineering, USICT, GGSIPU
Email: akansha.trar03@gmail.com
** Department of Computer Science and Engineering USICT, GGSIPU
Email: ashish@ipu.ac.in

heart causes this disease. The major three arteries are LAD (left anterior descending), LCX (left circumflex), and RCA (right coronary artery). This blockage causes due to the build-up of plaque or cholesterol on the inner walls of arteries. This build-up is known as Atherosclerosis. This build-up of the plaque reduces the blood supply to the heart because of which the heart does not get sufficient oxygen. It causes chest pain known as angina. If the plague is in a large quantity, then it cuts offthe blood supply to the heart leading to a heart attack.

CAD is now the leading cause of deaths worldwide. According to the World Health Organization (WHO), 17.9 million people die each year because of CVD, which is 31% of all the deaths occurring worldwide. More than 75% of deaths caused by CVD occur in low and middle-income countries and, 85% of these deaths occur due toheart attack, angina, and stroke [1]. Many studies showed the impact of CAD in different countries or regions. Itwas found that in North America and Europe [2], CAD is one of the major factors causing heart failure in more than 50% cases. In the study of Saudi Arabia [3], it was found that middle-aged people are more vulnerable to this disease. The American Heart Association [4] had estimated that there are 18.2 million Americans of Age more than 20 years who has CAD. The study done by European society of cardiology [5] has showed middle-income countries in Europe are the major sufferers of this disease. From the above-mentioned studies, it could be determined that CAD is a major problem worldwide, where the causes are different for this disease. Hence, it is necessary to find better solutions for the early prediction of this disease. There is an invasive method for diagnosing CAD, which is the best and moist accurate method known as Angiography. However, it has certain limitations such as being expensive and have some side effects. It led the researchers to find out some non-invasive techniques such as machine learning.

Different areas are using machine learning for a very long time. One such area is the medical field. Data mining is the technique for extracting hidden patterns and knowledge. To date, various algorithms and other feature selection techniques have been applied to multiple heart datasets [6, 7] to achieve optimized results. Some of theprevious studies even performed normalization, discretization to make a good quality dataset and hence earned the best accuracy. In our research, we have tried to better the previous results for diagnosing CAD and achievingthe best accuracy. The main contribution of this study is as follows:

- solving the problem of class skewness in the dataset
- to find a good quality feature set
- to achieve the best accuracy.

The problem of class skewness never considered in any of the previous studies on CAD datasets. Each of theseobjectives has been discussed and explained in the later sections.

The organization of the paper is as follows. Section 2 describes the research work done in the field of CAD diagnosis using different feature selection methods and classifiers. Section 3 describes the dataset "extended Z- Alizadeh Sani dataset" in detail. This section also describes the second aspect of CAD diagnosis. It further describes the analysis of the dataset in fact. Here, we have discussed the imbalance problem of our dataset. Section 4 describes the complete process of the proposed study, covering all of its subsections in detail. Section 5 describes all the results and discussion related to our research and, Section 6 concludes the proposed study, and section 7 guides for future work.

## RELATED WORK

In this area, a lot of research was already accomplished. The main aim of all the investigations was to achieve thehighest accuracy for predicting CAD and also providing the best features. The most popular datasets used in these studies were: Z-Alizadeh Sani dataset, the extension of the Z-Alizadeh Sani dataset, and the Cleveland dataset. In this section, we will see some of the works that were previously done on these datasets.

Youness Khourdifi et al. [8] have proposed the FCBF (fast correlation-Based feature selection) method to filter redundant attributes to improve the classification. They have used various supervised classification algorithms optimized by PSO (particle swarm optimization) and ACO (ant colony optimization). Using this model, they were able to achieve an accuracy of 99.65%.

Majid Ghonji Feshki et al. [9] have proposed an evolutionary algorithm of PSO and feed-forward neural networks. They have used the Cleveland dataset. By using feature ranking along with PSO, they were able to reduce 13 features to 8 optimized features. Using the neural network of feed-forward backpropagation on the selected attributes, they were able to get an accuracy of 91.94%. Hui Hu et al. [10] have designed a PSO-based generic classifier, which is an improved algorithm. They have implemented this algorithm on the platform Waikato Environment for Knowledge (WEKA). They have also compared this classifier with other classifiers on two different datasets. Results have shown that the proposed algorithm could get more effective classifying results. Simon Fong et al. [11] have proposed a collection of search methods that implement metaheuristics search known as swarm search. They have also added these searches to the

WEKA workbench. There were nine search methods: ant, bat, bee, cuckoo, elephant, firefly, flower, harmony, and Wolf Search Algorithm (WSA). We will see the performances of all these searches on our dataset in section 4. Lucija Brezocnik [12] has proposed using a C4.5 classifier with binary particle swarm optimization on different 11 datasets. He has shown that this combination outperforms the other classifiers, such as C4.5, Naïve Bayes, and Support Vector Machine (SVM). Roohallah et al. [13] have proposed a novel approach for predicting the stenotic arteries. This novel approach includes three classifiers. They have also extended the number of records from 303 to 500 and achieved an accuracy of 96.40%.

C. Beulah et al. [14] have investigated ensembling techniques for improving the accuracy of predicting heart disease. They have performed their study on the Cleveland dataset. They have also done a comparative analysis of different methods. They formed different sets of features by using feature selection techniques. With the help of ensemble techniques, they were able to get an increase in accuracy of 7%. Moloud Abdar et al. [15] have performed ten traditional algorithms and three types of SVM on the Z-Alizadeh Sani dataset. For increasing the performance of the classifiers, they used normalization in the preprocessing step and generic algorithm and PSO for feature selection. They also introduced a new optimization technique called N2Genetic optimizer. By using N2Genetic and NuSVM, they were able to get the accuracy of 93.08%. Alizadeh Sani, R. et al. [16] have attempted to achieve a high rate of accuracy for each artery separately. For feature selection, they used two approaches. The first method, where they used weight-by-SVM to find different features for each artery. The second method, they used weight-by-information gain to find out the standard features for all arteries. By applying SVM on the selected features, they achieved an accuracy of 86.14%, 83.17%, and 83.50% for LAD, LCX, and RCA, respectively.

## DATASET AND ITS ANALYSIS

For our study, we have taken the "extension of Z-Alizadeh Sani dataset" from the UCI dataset library [17]. As the name suggests, it is an extended version of the original Z-Alizadeh Sani dataset. This dataset contains 303 instances and 56 attributes. The new addition to the original dataset was the three major arteries whose blockage causes coronary artery disease. This further information has helped us in diagnosing CAD more accurately. This information was collected from the Shaheed Rajaei Cardiovascular medical and research center, Tehran, Iran. The features in this dataset are further categorized into four groups: Demographic features, Symptoms and examination, ECG, and laboratory

and echo features.

In this study, we have focused on the second aspect for diagnosing CAD, which is to predict stenosis in all threearteries. Prediction is performed for all three arteries because they play a vital role in diagnosing CAD. The presence of >50% stenosis in any of the arteries indicates the presence of CAD. Since we are working on the second aspect of diagnosing CAD, our focus will be on predicting whether any of the arteries is stenotic or not. As we have mentioned earlier, the blockage of one or more of the three arteries causes Coronary artery disease. For all three arteries (LAD, LCX, and RCA), there are two classes. One class represents those people who had CAD (216 people in this case), and another class represents those people who were healthy.

As we have discussed in section 1, we aim to improvise the result of diagnosing CAD; hence we have thoroughly analyzed this dataset. Since we are using data mining as a non-invasive method for prediction, we have focused on every aspect of data mining. Data mining helps in discovering knowledge and hidden patterns. It is an amalgamation of specific steps wherein each step plays an essential role in prediction. Each step of data mining plays a significant role. If we miss any of the steps, then we will not be able to achieve improved accuracy and good performance. The performance of classifiers highly depends upon the data on which it is working. Hence, itis crucial that before we apply any classifier, we have high quality and accurate data.

Here, we have our dataset as the raw dataset which needs to be checked for any inconsistency, missing values, and irrelevant data, etc. Our dataset does not contain any inconsistency or missing values; however, it is an imbalanced dataset. There is class skewness in the dataset. We have 303 instances, wherein 216 belongs to the majority class (people diagnosed with CAD), and 87 belongs to the minority class (healthy people). This imbalance problem misleads the classifiers, and therefore, results in incorrect output. Hence, this problem needsto be solved. To date, no one has addressed this problem in any CAD study. The solution to this problem is discussed in section 4.

## PROPOSED WORK

This study proposes a solution to the problem of class skewness. In section 5, we will see that the removal of the class skewness problem has not only provided us a good quality dataset but also helped in achieving the highestaccuracy. This study also proposes that metaphor searches are the best search methods in determining a good quality feature set. Fig. 1 shows the basic flowchart of the proposed work.

The flowchart shows that the medical dataset is first divided into a training set and testing set in a ratio of 9:1. Thetraining dataset is preprocessed using two supervised filters, namely the *Resample filter* and *SMOTE filter*. Thesefilters have helped in balancing the dataset. After this, various feature selection methods are applied. All these methods are compared to multiple supervised classifiers in terms of accuracy. These experiments have been performed on the Weka 3-9-3 tool.

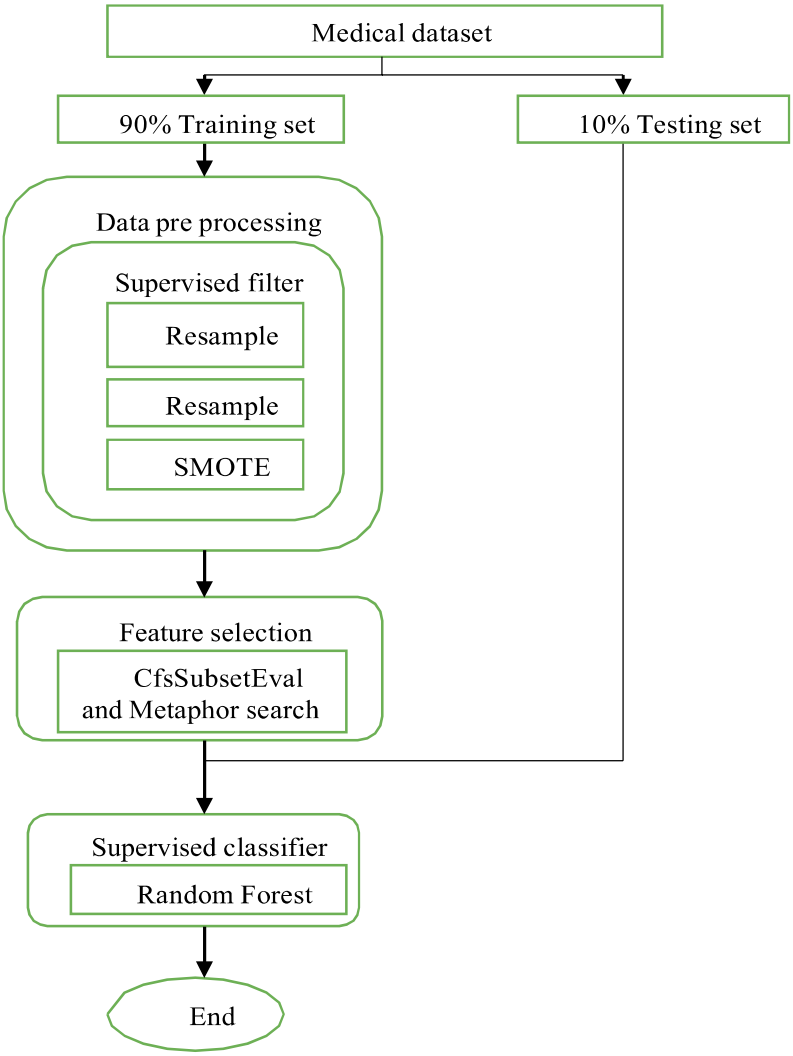



**Fig. 1: Flowchart of the proposed work**

All the stages of the proposed work are described below.

## Data preprocessing

As we have discussed in section 3, our dataset has a class skewness problem. To reduce this problem, we have used the *Resample* filter and, after that, the SMOTE filter.

The *Resample* filter is a supervised filter that produces a random subset of the dataset using sampling either with a replacement or without replacement. After this, the SMOTE filter is applied which is also a supervised filter.

SMOTE is Synthetic Minority Oversampling Technique. It has been used in various studies involving imbalanced datasets [18, 19]. It is one of the best techniques for imbalanced datasets. This technique acts only on the minority class. It increases the number of instances of the minority class by using the K-nearest neighbors. It does not make copies of the original instances. Rather it combines the features of the minority class and the chosen neighbors. In this way, it balances both the classes.

SMOTE algorithm is given below:

1. Choose a minority observation.
2. For each observation, find K nearest neighbors.
3. If the oversampling percentage is set at X%, let say 200%, then X (2) out of K neighbors will be selected randomly.
4. Create a synthetic sample over the line joining the observation and the selected neighbors.
5. Repeat this process until the dataset is balanced.

**Feature Selection**

Feature selection provides us those features that can help us in achieving the highest accuracy and improved performance. In Weka, feature selection consists of two parts: attribute evaluator and search methods. For different evaluation methods, multiple search methods are used. For attribute evaluators, we have chosen some of the filter approach methods. These methods are independent of the learning algorithm. They use the intrinsic properties of the data. These are *CfsSubsetEval, InfoGain, and GainRatio. CfsSubsetEval* [20] finds those subsets of features that are highly correlated with the target class but less correlated with other features. *InfoGain* and *GainRatio* use statistical measures to evaluate the worth of an attribute like entropy and gain ratio.

Some of the filter methods do not perform feature selection; they only rank them. That is why we go for search methods. These search methods, combined with these evaluators, provide the best features or subset of features. Weka includes local searches and evolutionary

type searches. Local searches are also known as conventional methods. These searches include the *best-first* search, *DepthFirst* search, etc. Evolutionary type searches are optimization techniques. These searches are iterative. They start with any random solution and then on each iteration, try to improvise the previous result. They keep on iterating until they cannot make any further improvements. Some examples of such techniques include PSO search, Genetic search, etc. There is also a different type of search (metaphor search), which was never considered in this type of study in the literature. Metaphor search is also known as the swarm search method. There are a total of nine search methods available inWeka. This search include *AntSearch, BatSearch, BeeSearch, CuckooSearch, ElephantSearch, FireflySearch, FlowerSearch, HarmonySearch, and WolfSearch.* These metaphor searches are metaheuristic optimization algorithms inspired by nature. Table I shows a brief description of these metaphor searches.

**Table I. A brief description of metaphor searches**

| Metaphor search | Description |
|---|---|
| Ant Search | The idea for these algorithms was generated from an analogy of ant colonies. |
| | Provides faster convergence |
| | Follows Greedy approach: chooses the path which was previously followed bythe maximum number of other ants |
| Bat Search | It is inspired by the echo-locating behavior of bats. |
| Bee Search | Follows a nature-inspired foraging activity of honey bees. |
| | Performs an explorative search and then an exploitative investigation among theselected search space. |
| | The waggle dance is considered as a fitness function. |
| Cuckoo Search | Based on obligate brood parasitism performed by the cuckoo. |
| | The aim is to find better solutions than the current ones and then replace them. |
| | New solutions are found using Levy flight. |
| Elephant Search | Based on the concept of behavioral characteristics of the elephant herds. |
| | It follows two concepts: the evolutionary concept and the dual balancing of exploratory search and exploitation search. |

| | |
|---|---|
| Firefly Search | It is inspired by the fireflies. |
| | Uses their bioluminescence property, which has two purposes: attractiveness andbrightness. |
| Flower Search | It is inspired by the pollination process in flowers. |
| | It involves two processes: global pollination which can perform Levy flight andlocal pollination when there are no pollinators. |
| | The task is to find the best solution either globally or locally. |
| Harmony Search | It is inspired by the artificial process of music performances. |
| | The algorithm will try to find out the best solution using an objective function. This objective function will be determined by the values of various componentsinvolved in the objective function. |
| Wolf Search | Bio-inspired by the preying behavior of wolves and the way they avoid theirenemies. |

All the previous studies show that most of the researchers used filter methods like InfoGain, Gini Index, weight-by-SVM, and different search methods. None of the studies performed the metaphor search. In our research, wehave performed feature selection using various combinations of attribute evaluation and search methods. Featureselection has been done for all the three arteries. Table II shows the number of features selected corresponding to different feature selection techniques.

**Table II. Number of features selected by different feature selection techniques for all the three arteries**

| **Feature selection techniques** | | **Number of s** | **elected features** | |
|---|---|---|---|---|
| Attribute evaluator | Search methods | RCA | LCX | LAD |
| | **Evolutionary type search** | | | |
| | Genetic | 23 | 20 | 23 |
| CFS | PSO | 15 | 17 | 21 |
| | Tabu | 16 | 16 | 14 |
| | **Metaphor search** | | | |
| | Ant | 10 | 9 | 16 |
| | Bat | 17 | 20 | 19 |
| | Bee | 23 | 16 | 15 |
| | Cuckoo | 19 | 17 | 23 |

| | | | | |
|---|---|---|---|---|
| CFS | Elephant | 20 | 17 | 24 |
| | Firefly | 17 | 22 | 19 |
| | Flower | 15 | 13 | 14 |
| | Harmony | 13 | 14 | 19 |
| | Wolf | 16 | 16 | 21 |
| | Local search best first | 16 | 16 | 17 |
| CFS | GreedyStepwise | 16 | 16 | 14 |
| | RankSearch | 22 | 29 | 38 |
| InfoGain | Ranker | 41 | 42 | 49 |
| GainRatio | Ranker | 41 | 42 | 44 |
| Correlation | Ranker | 54 | 30 | 54 |

**Supervised classifiers**

After preprocessing and feature selection was done, the selected number of features was given as input to different classifiers. These algorithms were then tested on these selected features using 10 fold cross-validations. The algorithms used in our study are discussed below.

- Random Forest

  It is an ensemble technique. It is used widely because of its simplicity, flexibility, and diversity (used for both classification and regression). It is an extension of the bagging technique. It is a collection of several decision trees. It avoids the problem of overfitting by choosing a random subset of features while splitting a node. It then creates multiple decision trees which are trained with bagging technique and then combine them. It makes them more accurate and provides a stable prediction.

- Bagging

  It is an ensemble technique. It combines multiple weak base learners to form a single strong learner. The idea is to train various models using the same learning algorithm. Different sample data is provided to the learners using sampling with replacement. When the test data is given to them, they generate different outputs. The final answer would be the one that is caused by the majority of the learners. It can be accomplished using a voting classifier.

- MLP

  It is an artificial neural network. It consists of multiple perceptrons. A perceptron is a linear classifier that classifies those datasets which are linearly separable. The structure of this complete neuron is complex. It consists of an input layer, a hidden layer, and an output layer. The number of the input layerand output layer is the same, but there can be any number of hidden layers depending upon the structure of the neural network.

  Each layer consists of several nodes or as we call them as neurons. Raw data is fed to the input layer. There are weights in between the input layer and the hidden layer. The dot product of the inputs and these weights determines the activation function for the hidden layer. The hidden layer performs the computation part. The output generated by the hidden layer is pushed towards the next layer using the activation function; it can either be a hidden layer or an output layer. The performance of the output layer can be determined by the weights between the hidden layer and the output layer.

- K-nearest neighbor

  It is one of the simplest supervised algorithms which is based on a similarity measure. It is mostly usedfor classification. It is used when we have labeled datasets which are noise-free and small because it isa lazy learner. The performance of any data point is determined by the performances of its nearest neighbors. The number of neighbors is determined by a parameter K. Choosing the right value of K is very important. In our study, we have taken K = 1, and the *nearestNeighborSearchAlgorithm* selected is linearNNSearch.

- J48

  It is a java implementation of the C4.5 algorithm in Weka. C4.5 is a software extension of the ID3 algorithm. The decision trees generated by the C4.5 algorithm are used for classification, and that is why they are known as a statistical classifier. They use the concept of information gain ratio while building decision trees. They generate decision trees from a set of classified samples. At each node, it chooses those attributes having the highest information gain ratio for making the decision. It can handle both continuous and discrete features.

## Performance Measures

Weka provides a lot of performance measures, some of which are very important in the medical field. For our study, we have used accuracy, sensitivity, and specificity measures. Based on these measures, we have compared the mentioned techniques. These measures are discussed below:

### (a) Accuracy

It defines the number of correct predictions made out of the total predictions. We can calculate accuracy using equation 1, where the parameters True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN) have been taken from the confusion matrix, as shown in Table III.

**Table III. Confusion matrix**

$$accuracy = (TP + TN)/(TP + TN + FP + FN), \quad (1)$$

Other measures that can also be determined from the confusion matrix are Precision and Recall, as shown in equation (2), and (3).

$$precision = TN/(TN + FP), \quad (2)$$

$$recall = TP/(TP + FN), \quad (3)$$

### (b) Kappa Statistics

Other measures that we have used are Kappa statistics, ROC (receiver operating characteristics) curve, and Attributes. Usually, accuracy is considered the primary performance measure, but it is not an effective measure. It does not take into account the imbalanced classification. Since our data was imbalanced, we have taken the kappa statistics measure. It is an effective measure that can handle the imbalanced class problem. It tells us how the classifier has performed across all instances. Equation 4 shows the formula for Kappa Statistics (K). O is the observed accuracy, and E is the expected accuracy.

$$k = (O - E)/(1 - E), \quad (4)$$

K's value can be between 0 and 1. Landis and Koch [**21**] have provided a way to characterize the values of k. On the basis of these values, we could easily determine, which classifier is best. These values are listed in Table IV.

**Table IV. Characterization of kappa statistic values**

| K's Values | Characteristics |
|---|---|
| 0-0.20 | Slight |
| 0.21-0.40 | Fair |
| 0.41-0.60 | Moderate |
| 0.61-0.80 | Substantial |
| 0.81-1 | Almost perfect agreement |

### (c) Receiver operating characteristics (ROC) curve

A ROC graph is a two-dimensional graph or curve. It is used for the visualization of the performance of a classifier. It helps in selecting the best classifier based on its performances. It depends upon two measures deducted from the confusion matrix. These two measures are TPR (true positive rate) and

FPR (false positive rate). TPR is also known as recall. The equations for TPR and FPR are shown in equation 5 and equation 6.

$$TPR = TP/(TP + FN), \quad (5)$$

$$FPR = FP/(FP + TN), \quad (6)$$

In a ROC graph, TPR forms the Y-axis, and FPR forms the X-axis. Y-axis represents the benefit, and X-axis represents a cost. The more the area under the curve, the higher will be the performance. The value of the ROC curve is between 0 and 1.

## RESULTS AND DISCUSSION

After performing preprocessing, feature selection, and running multiple classifiers on the different feature sets, we got the results, as shown in Table V, VI, and VII. As we have mentioned in earlier sections, we aim to improvise the previous results for predicting the CAD diagnosis. For that, we have performed preprocessing and feature selection. Both feature selection and preprocessing have a significant effect on accuracy, but there is a dilemma regarding which feature selection method is to be used. Different methods have provided different feature sets, and upon evaluation, they proved to be crucial features. As discussed in section 4, various feature selection methods were used for determining features. Here, we have compared feature selection methods for all

the three

arteries corresponding to different supervised classifiers mentioned in section 4. The comparison is shown inTable V, VI, and VII.

**Table V. Comparison of feature selection methods corresponding to some classifiers for LAD artery**

| **Feature selection method** | | **Algorithms** | | | | | |
|---|---|---|---|---|---|---|---|
| Attribute evaluator | Search method | RF | J48 | MLP | IBK | | Bagging |
| **Evolutionary search** | | | | | | | |
| CFS | Genetic | 96.27 | 93.93 | 94.63 | 94.17 | 96.03 | 92.07 |
| | PSO | 96.50 | 94.17 | 95.80 | 94.87 | | 92.54 |
| | Tabu | 96.73 | 95.10 | 94.63 | | | 93.24 |
| **Metaphor search** | | | | | | | |
| CFS | Ant | 96.03 | 93.70 | 94.40 | 96.27 | | 92.30 |
| | Bat | 96.73 | 95.33 | 94.17 | 96.27 | 95.33 | 93.24 |
| | Bee | 96.50 | 95.10 | 94.87 | 94.63 | | 94.40 |
| | Cuckoo | 97.20 | 94.87 | 95.33 | 94.63 | | 92.54 |
| | Elephant | 96.73 | 93.70 | 95.10 | 94.17 | | 91.60 |
| | Firefly | 96.03 | 94.63 | 95.57 | 95.33 | | 93.24 |
| | Flower | 95.80 | 94.63 | 94.40 | 95.57 | | 92.77 |
| | Harmony | 96.03 | 95.10 | 95.57 | 95.10 | | 92.77 |
| | Wolf | 95.57 | 94.63 | 95.33 | | | 92.30 |
| **Local search** | | | | | | | |
| CFS | best first | 96.73 | 94.63 | 93.24 | 95.57 | | 93.24 |
| | GreedyStepw | 96.73 | 95.10 | 94.63 | 94.87 | | 93.24 |
| | ise RankSearch | 96.96 | 95.57 | 95.10 | 95.10 | | 93.24 |
| InfoGain | Ranker | 96.96 | 94.63 | 94.17 | 94.40 | | 93.47 |
| GainRatio | Ranker | 97.20 | 95.10 | 94.87 | 93.93 | | 93.47 |
| Correlation | Ranker | 96.50 | 95.57 | 94.40 | 93.70 | | 93.47 |

**Table VI. Comparison of feature selection methods corresponding to some classifiers for LCX artery**

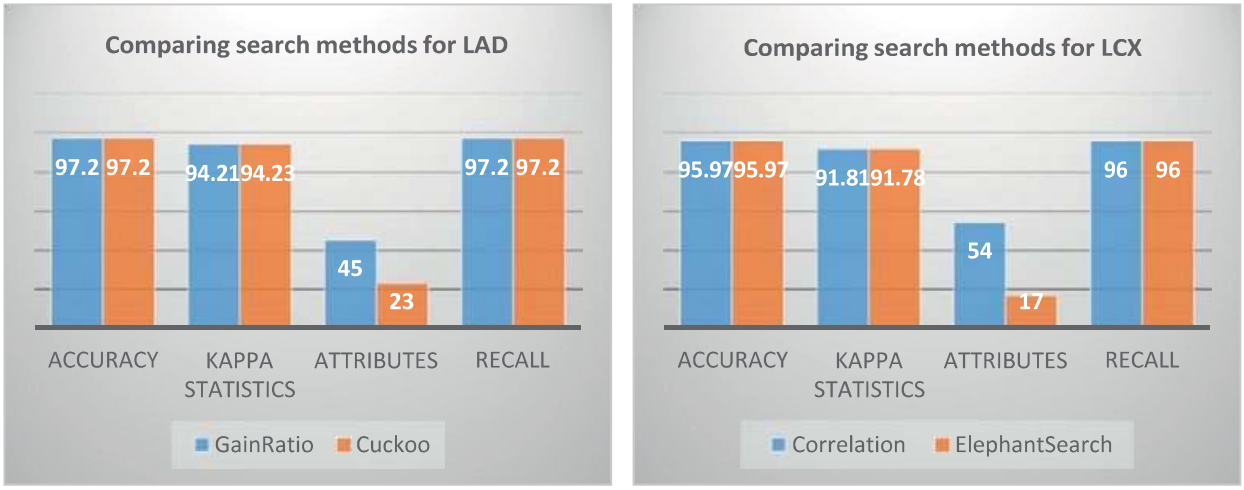
| Feature selection method | | Algorithms | | | | |
|---|---|---|---|---|---|---|
| **Attribute evaluator** | **Search method** | **RF** | **J48** | **MLP** | **IBK** | **Bagging** |
| **Evolutionary search** | | | | | | |
| CFS | Genetic | 94.78 | 89.33 | 93.60 | 93.83 | 90.28 |
| | PSO | 95.26 | 88.62 | 91.70 | 94.54 | 90.04 |
| | Tabu | 95.02 | 90.04 | 92.89 | 94.31 | 90.28 |
| **Metaphor search** | | | | | | |
| CFS | Ant | 95.26 | 92.89 | 87.91 | 92.65 | 90.99 |
| | Bat | 95.26 | 90.28 | 91.46 | 92.89 | 90.99 |
| | Bee | 95.97 | 89.81 | 91.46 | 92.65 | 89.81 |
| | Cuckoo | 95.26 | 88.62 | 91.70 | 92.89 | 90.28 |
| | Elephant | 95.97 | 91.46 | 90.75 | 93.36 | 90.04 |
| | Firefly | 95.73 | 90.75 | 92.89 | 91.70 | 90.52 |
| | Flower | 95.73 | 90.04 | 91.23 | 93.36 | 91.23 |
| | Harmony | 94.78 | 89.81 | 91.23 | 93.60 | 88.15 |
| | Wolf | 95.02 | 90.04 | 92.89 | 94.31 | 90.28 |
| **Local search** | | | | | | |
| CFS | best first | 95.02 | 90.04 | 92.89 | 94.31 | 90.28 |
| | GreedyStepwise | 95.02 | 90.04 | 92.89 | 94.31 | 90.28 |
| | RankSearch | 94.78 | 90.52 | 92.18 | 93.12 | 90.28 |
| InfoGain | Ranker | 95.26 | 89.57 | 94.07 | 93.83 | 90.75 |
| GainRatio | Ranker | 95.26 | 89.57 | 94.07 | 93.83 | 90.75 |
| Correlation | Ranker | 95.97 | 89.81 | 93.36 | 94.78 | 89.57 |

**Table VII. Comparison of feature selection methods corresponding to some classifiers for RCA artery**

| Feature selection method | | Algorithms | | | | |
|---|---|---|---|---|---|---|
| Attribute evaluator Search method | RF | J48 | MLP | IBK | Bagging | |
| **Evolutionary search** | | | | | | |
| CFS | Genetic | 93.76 | 87.76 | 91.12 | 92.56 | 87.05 |
| | PSO | 93.52 | 90.16 | 91.84 | 91.60 | 88.24 |
| | Tabu | 93.28 | 88.72 | 88.48 | 91.12 | 89.68 |
| **Metaphor search** | | | | | | |
| CFS | Ant | 94.48 | 90.64 | 88.24 | 90.64 | 87.29 |
| | Bat | 93.76 | 89.92 | 88.96 | 91.60 | 89.20 |
| | Bee | 93.28 | 87.76 | 93.52 | 93.76 | 87.05 |
| | Cuckoo | 93.76 | 89.92 | 87.29 | 91.36 | 89.44 |
| | Elephant | 92.56 | 89.68 | 89.68 | 91.12 | 87.29 |
| | Firefly | 92.56 | 88.96 | 88.00 | 91.12 | 89.68 |
| | Flower | 92.80 | 89.92 | 89.44 | 91.36 | 88.00 |
| | Harmony | 93.52 | 88.48 | 89.68 | 90.40 | 89.20 |
| | Wolf | 93.28 | 88.72 | 88.72 | 91.12 | 89.68 |
| **Local search** | | | | | | |
| CFS | best first | 93.28 | 88.72 | 88.48 | 91.12 | 89.68 |
| | GreedyStepwise | 93.28 | 88.72 | 88.48 | 91.12 | 89.68 |
| | RankSearch | 93.28 | 88.96 | 92.56 | 90.64 | 88.00 |
| InfoGain | Ranker | 93.76 | 88.72 | 93.76 | 91.60 | 88.96 |
| GainRatio | Ranker | 93.76 | 88.72 | 93.76 | 91.60 | 88.96 |
| Correlation | Ranker | 94.00 | 90.40 | 92.08 | 93.04 | 91.36 |

Tables V, VI, and VII show that for all the feature selection methods, the random forest has achieved the highestaccuracy as compared to other methods. The accuracies achieved using different feature selection methods are nearly similar. However, if we look closely in Table V, then we shall see that *GainRatio*+Ranker and Cfs+*CuckooSearch* with random forest classifier has achieved the highest accuracy (97.20%) for the LAD artery. Similarly, in Table VI, Correlation+Ranker and Cfs+ElephantSearch with random forest have achieved the highest accuracy (95.97%) for LCX accuracy. For the RCA artery, Cfs+AntSearch, with the random forest, has achievedthe highest accuracy (94.48).

In Fig. 2(a), a comparison between *GainRatio*+Ranker and Cfs+*CuckooSearch* has been made, and in Fig. 2(b), a comparison between Correlation+Ranker and Cfs+ElephantSearch has been made. The comparison has been done in terms of accuracy, Kappa statistics, number of attributes, and recall. These comparisons give us the best method for achieving the best results.




*(a)* *(b)*

**Figure 2: part (a) shows the comparison of the GainRatio and Cuckoo method for the LAD artery. Part (b) shows the comparison of the Correlation and ElephantSearch method for the LCX artery.**

After comparing the performances of both the methods for LAD and the LCX artery as shown in Figure 1, we cansee that the performance of both the methods is nearly the same except for the number of attributes in *GainRatio*.Also, the number of attributes in *Correlation* are more than the number of attributes in metaphor search methods. With the average number of attributes, a metaphor search is also giving us the best results. Also, *Cfs+metaphor* searches can be considered better than *GainRatio+ranker* and *Correlation+ranker* method because the latter filter methods only rank the attributes

based on their statistical measures. They include redundant attributes, i.e., attributes having the same values. In contrast, the *Cfs+metaphor* search provides us a subset of irredundant features. Hence, metaphor searches are considered better than local searches.

We have a SMOTE filter for preprocessing and metaphor searches as feature selection methods. An accuracy comparison has also been done among different supervised classifiers for the three arteries. We ran these classifiers on the feature set obtained from metaphor searches for each artery. The comparison can be seen in Fig.3.

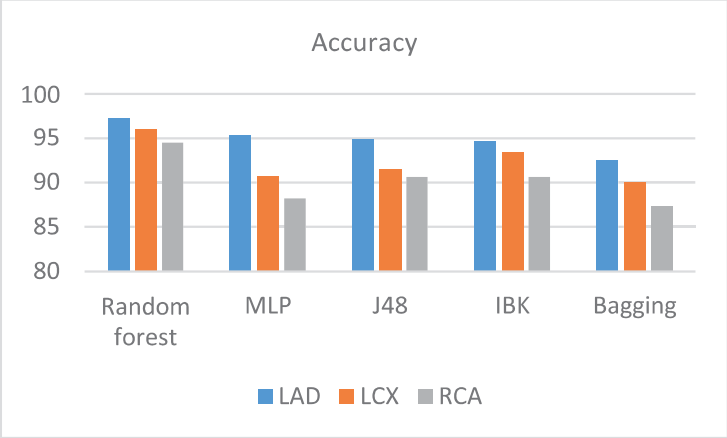



**Figure 3. Accuracy comparison among supervised classifiers for all arteries.**

Fig.3 shows that for all the three arteries, Random forest has achieved the highest accuracy. The second-best performance has been obtained by the IBK classifier. A ROC curve comparison among all three arteries can alsobe seen in Fig. 4.

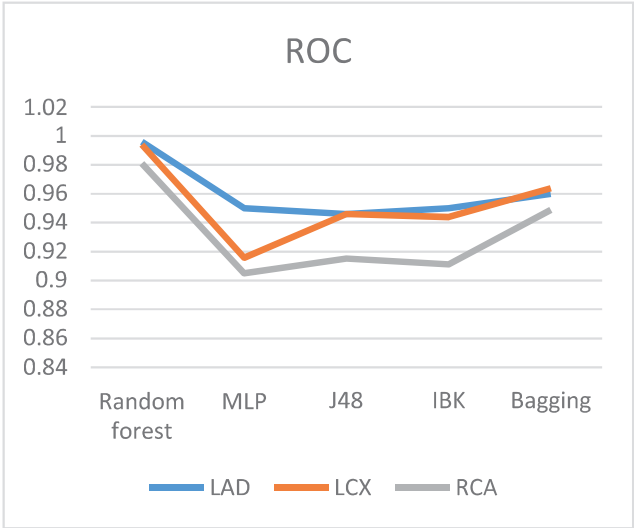



**Figure 4. ROC curve comparison among supervised classifiers for all the three arteries**

The higher the curve, the better will be the performance. Among all the three arteries, the LAD artery has achieved the best performance.

So, with SMOTE as a preprocessing method and Metaphor Searches as a feature selection method with random forest classifiers, we can achieve the highest accuracy of 97.20%, 95.97%, and 94.48% for artery LAD, LCX, andRCA respectively. To the best of our knowledge, these accuracies are the highest in the literature.

As we have seen in Table II, different feature selection methods have provided a different number of features. Some features were selected by almost all of the methods, and some were not. So after performing a manual analysis, we have found out some attributes are very important in diagnosing CAD because they were selected byall the methods. In our study, we have taken 18 methods for feature selection, and we have considered those features as "essential" which were selected by all the 18 methods.

The topmost important features for LAD are shown as:

| Age, typical CP, FBS, LDL, WBC > function class, PLT > weight, HDL, BUN, HB, Neut, EF-TTE |
|---|

Similarly, for LCX are:

| Age, BP, typical CP, TG, BUN > BMI, HTN, Atypical, PLT > length, WBC, lymph |
|---|

For RCA are:

| WBC, typical CP, HDL > DM, DLP, LDL, Age. |
|---|

We also found out that there is one feature that is never considered by any technique. It is Exertional CP. So wecan skip this feature. Among LAD, LCX, and RCA, the common features are **Age, typical CP, and WBC**. So, according to our proposed work, all the mentioned attributes in the boxes are essential in diagnosing CAD.

## CONCLUSION

The proposed study aims to improvise the previous works on predicting stenosis in coronary arteries. The proposed study has focused on three main issues to achieve this goal. The first issue was to tackle down the problem of class skewness. The proposed study has preferred the SMOTE filter to solve this problem. It is of utmost importance to have a good quality dataset before performing prediction. Otherwise, the chances of

getting biasedresults would be very high.

The second issue was to provide a final set of features for the early prediction of CAD. This study has compared eighteen different feature selection techniques so that the spectrum of research could get widened. Metaphor search (swarm search) has also been included because they were never considered in such type of study. These searches have been compared concerning different supervised classifiers. After performing all experiments, the study concludes that the random forest classifier and the metaphor search has outperformed the other techniques.The reason being that metaphor searches have provided an irredundant feature set, and they are giving the best performance even with a smaller number of features as compared to other techniques. The study also concludes that Age, typical chest pain, WBC, FBS, HDL, LDL, BP are some of the essential features.

The third issue was to achieve the highest accuracy in the literature. Achieving the highest accuracy depends upon several factors like dataset quality, feature set, and the classifiers used for the experiment. With the help of the SMOTE filter and Metaphor search, we could get a good quality feature set. The highest accuracy obtained for LAD, LCX, and RCA is 97.2%, 95.97%, and 94.48%, respectively. It can also be concluded that among all the three arteries, LAD has achieved the highest accuracy.

## FUTURE WORK

In the future, we would aim to find out the percentage of stenosis present in the arteries. In this study, we have considered an Iranian dataset. So, in the future, we may consider datasets from different countries and try to findout the causes of CAD distinctively, which could, in turn, help in early prediction of this.

**Acknowledgment**

None

Funding source

No funding was received for conducting this study.

**Conflict of interest**

The authors have no conflicts of interest to declare that are relevant to the content of this article.Ethics approval

Consent to participate

Consent for publication

Availability of data and material

## REFERENCES

Cardiovascular Disease (CVDs), [accessed on 15/11/2019] https://www.who.int/news-room/fact- sheets/detail/cardiovascular-diseases-(cvds).

Pagliaro, B.R., Cannata, F., Stefanini, G.G., and Bolognese, L. "Myocardial ischemia and coronary diseasein heart failure". *Heart Failure Reviews*, *25*(1), pp.53-65, 2020.

Albugami, S., Al-Husayni, F., Bakhsh, L., Alhameed, F., Alsulami, A., Abumelha, K., Balubaid, M., Al- Harbi, M. and Mufti, H.N. "The Perception of Coronary Artery Disease and Cardiac Catheterization in Saudi Arabia:What the Public Know". *Cureus*, *12*(1), 2020.

Virani, S.S., Alonso, A., Benjamin, E.J., Bittencourt, M.S., Callaway, C.W., Carson, A.P., Chamberlain, A.M., Chang, A.R., Cheng, S., Delling, F.N. and Djousse, L. Heart Disease and Stroke Statistics-2020 Update: A Report From the American Heart Association. *Circulation*, p.CIR0000000000000757, 2020.

Timmis, A., Townsend, N., Gale, C.P., Torbica, A., Lettino, M., Petersen, S.E., Mossialos, E.A., Maggioni, A.P., Kazakiewicz, D., May, H.T. and De Smedt, D. European Society of Cardiology: cardiovascular disease statistics 2019. *European Heart Journal*, *41*(1), pp.12-85, 2020.

Govindarajan, P., Soundarapandian, R.K., Gandomi, A.H., Patan, R., Jayaraman, P. and Manikandan, R. "Classification of stroke disease using machine learning algorithms". *Neural Computing and Applications*, *32*(3), pp.817-828, 2020.

Joloudari, J.H., Hassannataj Joloudari, E., Saadatfar, H., GhasemiGol, M., Razavi, S.M., Mosavi, A., Nabipour, N., Shamshirband, S. and Nadai, L. "Coronary artery disease diagnosis; ranking the significant features using a random trees model". *International journal of environmental research and public health*, *17*(3), p.731, 2020.

Khourdifi, Y. and Bahaj, M. "Heart Disease Prediction and Classification Using Machine Learning Algorithms Optimized by Particle Swarm Optimization and Ant Colony Optimization". *International Journal of Intelligent Engineering & Systems*, *12*(1), 2019.

Feshki, M.G. and Shijani, O.S., April. "Improving the heart disease diagnosis by evolutionary algorithm of PSO and Feed Forward Neural Network". In *2016 Artificial Intelligence and Robotics (IRANOPEN)* (pp. 48-53). IEEE.

Hu, H., Mao, X. and Xi, Q., 2017, March. "A PSO-based Generic Classifier Design and Weka Implementation Study". In *2016 International Forum on Mechanical, Control and Automation (IFMCA 2016)*. Atlantis Press.

Fong, S., Biuk-Aghai, R.P. and Millham, R.C. "Swarm search methods in weka for data mining". In *Proceedings of the 2018 10th International Conference on Machine Learning and Computing* (pp. 122- 127), 2018, February.

Brezočnik, L. "Feature selection for classification using particle swarm optimization". In *IEEE EUROCON2017-17th International Conference on Smart Technologies* (pp. 966-971). IEEE, 2017, July.

Alizadehsani, R., Hosseini, M.J., Khosravi, A., Khozeimeh, F., Roshanzamir, M., Sarrafzadegan, N. and Nahavandi, S.. "Non-invasive detection of coronary artery disease in highrisk patients based on the stenosis prediction of separate coronary arteries". Computer methods and programs in biomedicine, 162, (2018) pp.119-127.

Latha, C.B.C. and Jeeva, S.C. "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques". *Informatics in Medicine Unlocked*, *16*, p.100203, 2019.

Abdar, M., Książek, W., Acharya, U.R., Tan, R.S., Makarenkov, V. and Pławiak, P. "A new machine learning technique for an accurate diagnosis of coronary artery disease". *Computer methods and programsin biomedicine*, *179*, p.104992, 2019.

Alizadehsani, R., Zangooei, M.H., Hosseini, M.J., Habibi, J., Khosravi, A., Roshanzamir, M., Khozeimeh,F., Sarrafzadegan, N. and Nahavandi, S. "Coronary artery disease detection using computational intelligence methods". *Knowledge-Based Systems*, *109*, pp.187-197, 2016.

Extention of Z-Alizadeh sani dataset Data Set, https://archive.ics.uci.edu/ml/datasets/extention

+ of+Z-Alizadeh+sani+dataset. [accessed on February 15, 2020].

Sun, J., Li, H., Fujita, H., Fu, B. and Ai, W. "Class-imbalanced dynamic financial distress prediction based on Adaboost-SVM ensemble combined with SMOTE and time weighting". *Information Fusion*, *54*, pp.128-144, 2020.

Abdoh, S.F., Rizka, M.A. and Maghraby, F.A. "Cervical cancer diagnosis using random forest classifier with SMOTE and feature reduction techniques". *IEEE Access*, *6*, pp.59475-59485, 2018.

Class CfsSubsetEval, [accesses on 15/03/2020] https://weka.sourceforge.io/doc.dev/weka/attributeSelection/CfsSubsetEval.html

Landis, J.R. and Koch, G.G. "The measurement of observer agreement for categorical data". *biometrics*, pp.159-174, 1977.