ANALYSIS OF CARDIAC ANOMALIES BY SELECTION AND EXTRACTION OF FEATURES USING MACHINE LEARNING METHODS

S. Das

Research Scholar, Dept. of Computer Science and Applications, Utkal University,
Bhubaneswar, Odisha, India
subhalaxmi.das@gmail.com

S. K. Pradhan

Former Prof., Dept. of Computer Science and Applications, Utkal University,
Bhubaneswar, Odisha, India
sateesh1960@gmail.com

S. Mishra

Asst. Prof., Department of Mathematics, Odisha University of Technology and Research,
Bhubaneswar, Odisha, India
sujogya123@gmail.com

S. Pradhan

Asst. Prof., Department of Computer Science RBVRR Women's College,
Hyderabad, Telangana, India
sipalipradhan06@gmail.com

P. K. Pattnaik

Asst. Prof., Department of Mathematics, Odisha University of Technology and Research,
Bhubaneswar, Odisha, India
papun.pattnaik@gmail.com

Abstract: One of the main factors that contribute to a person's death throughout the world is cardiac disease nowadays. Hospitals have vast amounts of clinical data stored in biomedical devices and other systems. Understanding the facts and finding the most significant features that can increase the estimate accuracy that may cause heart disease is crucial. In this paper, a model has been proposed by using relevant characteristics of selected various feature selection, feature extraction, and machine learning classifiers. Filter and wrapper methods are used for feature selection, whereas principal component analysis (PCA) and linear discriminant analysis (LDA) are used for feature extraction. Both approaches were applied to the Cleveland heart disease dataset to get the relevant attribute subset. Once the attribute subset is selected, different machine learning classifiers, namely Logistic Regression (LR), Naïve Bayes (NB), Random Forest (RF), Extreme Gradient Boost (XGB), K-Nearest Neighbour (KNN), Decision Tree (DT), and Support Vector Classifier (SVC) were used to classify the presence and absence of heart disease. Finally, the top three classifiers with the most significant attributes and different feature selection/extraction techniques are discussed. The performance of the proposed model was also calculated using various performance evaluation metrics like accuracy, precision, recall and f1 score.

Keywords: PCA; LDA; Filter method; Wrapper Method; Accuracy; Precision

1. Introduction

One of the primary causes of death in the modern era is cardiovascular disease. The increased collection of medical data gives clinicians a new possibility to improve the patient's diagnosis. Experts have recently improved their use of computer technology to help them make better decision. Heart disease is a group of disorders that affect the heart, arteries, muscles, valves, and internal electrical circuits that control muscle contraction. Heart disease is the main cause of mortality in India, the United Kingdom, the United States, Canada, and Australia, according to the "Centers for Disease Control and Prevention". Around the world, cardiovascular diseases (CVDs) account for 31% (17.9 million) of all fatalities annually and are a significant contributor to clinical (death and disability), health, and economic burden. Heart disease accounts for one out of every four fatalities in the United States [2]. In most countries, both men and women are affected by heart disease. As a result, people should consider the risk factors for cardiovascular disease. Although it has a genetic component, certain lifestyle variables substantially impact heart disease. Malnutrition, high blood pressure, high blood cholesterol levels, diabetes, obesity, physical mobility, anxiety, and poor hygiene are all known risk factors for heart disease [3]. These are the different risk factors that increase a patient's chances of having CVD. Practitioners have a difficult time making a diagnosis since the symptoms are similar to those of other conditions or might be mistaken for signs of ageing.

Traditional statistical methods-based predictive models typically only capture some of the complicated causal linkages between multiple risk factors [4-5]. During a pandemic, Patients can use the Internet of Things (IoT) to feed information into a database by wearing smart wristbands and medications that can monitor and collect specific information [6]. Researchers use artificial intelligence (AI) to mine fresh medical data to assist physicians in analysing and understanding symptoms, making early predictions of illnesses, and making decisions for a variety of patient scenarios, effectively advancing the development of safe diagnostic tools [7]. The present systematisation of national health screening data and the standardisation of clinical data have made it possible to analyse risk variables that weren't previously known. [8-10]. These factors might be statistically substantially linked to the development of diseases which further allowing us to trace numerous disease pathways. Furthermore, big data analysis is critical for developing effective illness prediction models [11,12]. Machine learning has recently received great attention and is being utilised to predict CVD models.

Machine learning is quickly becoming a viable option in healthcare for assisting with patient diagnostics. When a dataset is vast and challenging to programme, such as when transforming a clinical record into knowledge, predicting pandemics, or analysing genetic data, machine learning like predictive model is applied.

Some are useless when a dataset has various features and causes poor outcomes. When using a computer-based system to diagnose different diseases, high-dimensional and diverse data must be managed and processed, and because of this high-dimensional data, the overfitting and training time of the model also increases. One method for reducing the quantity of data features into a manageable collection is feature selection, which works by eliminating superfluous, pointless, and redundant traits that have no bearing on the performance of the classifier. To combat the dimensionality curse, many efficient feature selection techniques have recently been created. Therefore, the main objective of this research is to use a combination technique to enhance feature selection and classification, resulting in a better diagnosis of heart disease. In this paper, we have used filter and wrapper method features to maximise selecting relevant features in heart disease, and then different classification techniques are utilised.

The residual of the paper is structured as follows: section 2, which contains related work; section 3, which covers the methodology utilised in this study; section 4, which contains the planned work; and section 5, which has the findings and related discussion. Section 6 concludes by outlining the conclusions and upcoming projects.

2. Related Work

Machine learning has a lot of difficulties when dealing with large dimensional datasets [12]. Weighting features decrease redundant data and processing time, which improves algorithm performance since the analysis of many features consumes a lot of memory and leads to overfitting [13–17]. Disorders in health management, genomic expression, medical images, and IoT are characterised by a small set of characteristics. Dimension reduction uses feature extraction to change and simplify data, whereas feature selection reduces the dataset by eliminating redundant traits [18]. The use of feature selection strategies to forecast heart disease was shown to be more accurate in the literature.

Ritu et al. [13] developed a sequential attribute selection technique that would be able to pinpoint the most crucial factors and detect the mortality events in heart disease patients throughout therapy. They have used several machine learning methods like linear discriminant analysis, k-nearest neighbor, random forest, decision tree, and support vector machine, and the performance of the model was evaluated using performance metrics. It showed that the result obtained by random forest is 86.67%.

Gao et al. [14] proposed a model for predicting cardiac disease. They have used the ensemble method and feature extraction methods, and additionally, the Cleveland heart disease dataset was utilised. The research' findings indicated that the bagging ensemble learning approach with DT and PCA produced superior outcomes to others.

Takci [15] employed feature selection methods and classification algorithms from distinct categories to predict heart attacks. The models' accuracy, processing speed, and ROC analysis results were assessed. The outcome demonstrates that the maximum accuracy value was 82.59% without feature selection and 84.81% with feature selection. Naive Bayes and linear SVM were used to achieve a model accuracy of 84.81%. The ReliefF algorithm delivers the optimised results. A feature selection and ensemble classification technique were put out by Latha et al. [16] for the prediction of the risk of heart disease. The results of the study showed that ensemble approaches, such as bagging and boosting, may accurately forecast the risk of acquiring heart disease and enhance the prediction accuracy of poor learners. Chi-square and principal component analysis were used to create a hybrid dimensionality reduction approach by Karen Garate-Escamila et al. [17]. They have chosen Hungarian, Cleveland, and Hungarian-Cleveland datasets for their work. Random forests, gradient-boosted trees, decision trees, multilayer perceptron, and logistic regression were used for further classification. With 98.7% accuracy for the Cleveland dataset, 99.0% accuracy for the Hungarian dataset, and 99.4% accuracy for the Cleveland-Hungarian (CH) dataset, chi-square and principal component analysis using random forests (RF) demonstrated the best accuracy. Principal component analysis, Chi-squared testing, ReliefF, and symmetrical uncertainty were used to suggest model which is authored by Spencer et al. [18]. The authors also discussed the advantages of selecting features used for the cardiac datasets. Semen et al. [19] constructed a method for diagnosing chronic renal illness. Recursive feature elimination was used to determine which characteristics were most important. K-nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree (DT), and Random Forest were employed in this study as classification techniques. The model's effectiveness was also assessed, and they found that random forest achieved 100% accuracy. Almansour et al. [20] presented a model that can predict chronic kidney disease. They used the “UCI repository’s 400 instance dataset” for their research. Support vector machines and artificial neural networks were used as classifiers. The effectiveness of the classifiers is assessed using the best features and by calculating the model's training duration.

The body of research demonstrates that classifier accuracy is increased by training them with relevant characteristics selected by various feature selection and feature extraction techniques.

3. Methodology

This section will provide ideas about the different methodologies that are used in this research work. The basics of feature selection and extraction techniques are described in the following sections.

3.1 Features Selection

Picking the right characteristics to attain the best classification outcome for the statistics is one of the most challenging challenges in the classification issue. Additionally, experimental data demonstrates that each attribute is required for correctly identifying the data class label. With the use of feature selection, models may be made simpler and more computationally efficient by selecting a subset of their original attributes. Filtering, wrapping, and embedding feature selection techniques are the three different types [21-23].

3.1.1. Filtering methods

Filtering processes employ indirect criteria to evaluate the accuracy of predictions or classifications, such as the distance measure, which illustrates how effectively the attributes are parted. Usually, a pre-processing phase is performed using this technique.

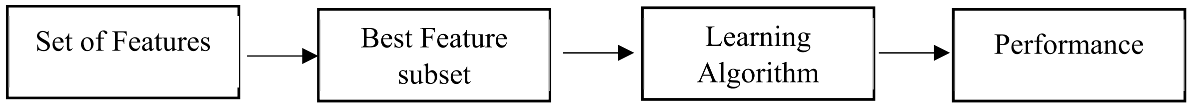

Fig 1. Steps in Filter Method

As shown in Figure 1, in the filter methods, various statistical tests are used to associate the characteristics, and those features are chosen based on their computed scores or weight values [24-32]. Different filtering technologies exist, some of which are described below.

- **Fisher Score**

The Fisher Score is a technique for identifying the subset of features or attributes most crucial for classification problems. It uses generative statistical models and discriminative approaches to get the optimal features. The characteristics are chosen based on the rule that distances between attribute points fitting to the same cluster should be minimal, and those corresponding to other clusters should be substantial. According to their scores under the Fisher criterion, each feature is individually chosen by Fisher's Score, resulting in a suboptimal collection of features. Generally, the feature having higher fisher scores are selected.

For binary classification, the fisher score (FS) of the ith feature is defined as

$$\mathrm{FS}=\frac{\overline{X}_i^{(0)}-\overline{X}_i^{(1)}}{\sqrt{\mathrm{var}(X_i)^{(0)}+\mathrm{var}(X_i)^{(1)}}} \tag{2.1}$$

where$\overline{X}_i^{(0)}$, $\overline{X}_i^{(1)}$,var $(X_i)^{(0)}$ and var $(X_i)^{(1)}$, are the means of sample and variances of feature i, for the recurring patterns in each class.

- **Variance Threshold**

In Variance Threshold, it eliminates all low variance characteristics which are not useful in the dataset. So, it can be applied to unsupervised learning since it only considers the characteristics (x), not the expected or target outputs (y). Only numerical data is suitable for the Variance Threshold. This approach eliminates characteristics with zero variance by default, and this technique assumes that characteristics with larger variations are more likely to contain information.

- **Mean absolute difference (MAD)**

This approach is comparable to the variance threshold method, except MAD does not use squares. This approach determines the mean absolute deviation from the mean. The mean absolute difference (MAD) is defined as

$$\mathrm{MAD}_i=\frac{1}{n}\sum_{j=1}^{n}|X_{ij}-\overline{X}_i| \tag{2.2}$$

Determines the absolute deviation from the mean.

- **Dispersion ratio**

The ratio of the geometric mean (GM) to the arithmetic mean (AM) for a particular feature is known as the dispersion ratio. For a specific characteristic, it has a value that goes from +1 to GM as AM. A higher dispersion ratio suggests a more important trait.

- **Exhaustive Feature Elimination**

In wide feature selection, every possible combination of the features in the data frame is used to assess how well a machine learning algorithm performs. Since it tests every possible combination of characteristics, the exhaustive search algorithm picks the best and the most avaricious of all the wrapping strategies.

- **Correlation**

The Correlation technique can be used to find the association between variables. If the features are strongly correlate with the target, then the correlation may be used to select features. Variables should also be uncorrelated among themselves while being associated with the target. We can anticipate one variable from another if the two are associated. So, if two characteristics are linked, the model only requires one feature as the other does not provide any new information. The degree to which two variables are linearly connected is expressed by a Pearson correlation, which ranges from -1 to 1.

- **Mutual Information Gain**

It can be performed on classification and regression. It works on the entropy of the variables. Positive values indicate that two random variables are interdependent, as measured by mutual information (MI). Greater dependency is shown by bigger numbers; dependence equals 0 only when two random variables are independent. The following definition identifies the mutual information between two random variables, A and B:

$$I(A;B)=H(A)-H(A|B) \tag{2.3}$$

Where I (A; B) is the mutual information for random variables A and B, H(A) is the entropy for A, and H(A|B) is the conditional entropy for A given B. The value of the result lies between zero and one.

- **Chi-Square**

In general, the categorical properties of a dataset are tested using the Chi-square method. Chi-square is used to analyse each feature's relevance to the target, and the desired number of attributes are then choosen according to the Chi-square scores. The dataset's observed values for several properties are contrasted with the expected value.

$$X^2=\sum\frac{(\text{Observed value}-\text{Expected value})^2}{\text{Expected value}} \tag{2.4}$$

3.1.2. Wrapper methods

In wrapper approaches a subset of features throughout the different search phases using a search strategy and a learning mode, as shown in Fig. 2. Wrapper techniques often outperform filter methods in classification accuracy due to using a learning model. On the other hand, they have several drawbacks, including a large computational overhead and the potential for overfitting. Examples of common wrapper techniques are forward feature selection, backward feature removal, bidirectional feature elimination, and exhaustive feature selection.

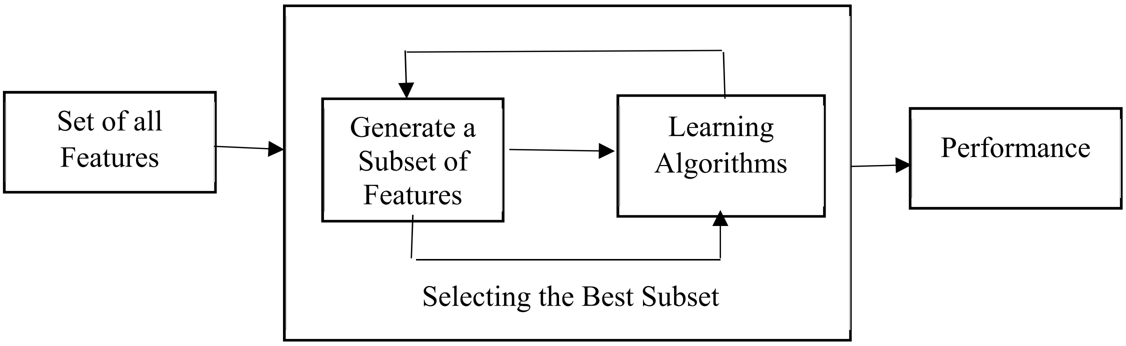

Fig. 2 Steps in Wrapper Methods

- **Forward feature selection**

It is a continuous process. The selecting process begins with a model devoid of any characteristics. Every iteration adds a new feature up until the point where the performance matrices of the model are unaffected by the adding of a new variable.

- **Backward feature removal**

Backward elimination helps the model perform better by starting with all the features and removing the least important one at a time. We keep doing this until we see no improvement when we remove features.

- **Bidirectional feature elimination (Step-wise selection)**

Similar to the forward selection, it also considers the applicability of recently additional features whenever a new attribute is induced. If any previously selected attributes are irrelevant, backwards selection simply eliminates that feature. As a result, it combines backward elimination with forward selection.

- **Exhaustive feature selection**

In this selection process, every possible combination of the dataset's characteristics is tested against the performance of a machine learning algorithm. The attribute subset with the best outcome is chosen. Since it tests every combination of characteristics and chooses the most efficient combination of the feature set, the exhaustive search algorithm is the most greedy among all of the wrapper approaches. Since it considers all attribute amalgamations, it could be more time consuming than the forward selection and backward selection technique.

3.1.3 Embedded Methods

In the learning process, these techniques choose features, and learners are typically given these tasks. This model benefits from the strengths of the other models by using various assessment standards at various phases of the search process. Wrapper and filter functions are combined in embedded techniques. These techniques include built-in penalisation procedures to reduce overfitting, including some of the most well-known examples, LASSO and RIDGE regression. Lasso regression employs L1 regularisation, which introduces a penalty proportional to the absolute value of the degree of the coefficients. The penalty for L2 regularization, which is used in ridge regression, is inversely correlated with the square of the size of the coefficients.

3.2. Feature Extraction

The motive of this feature extraction technique is to build a new feature subspace by extracting or deriving data from the original feature set. Feature extraction's main purpose is to keep the most pertinent information while compressing the data. Similar to the feature selection approaches, these strategies are also replicate to minimise the count of attributes from the initial set of attributes to decrease model complexity, reduce model overfitting, increase the efficiency of model computation, and decrease generalisation error. There are several feature extraction methods, including the following.

3.2.1. Principal component analysis

A technique for decreasing the dimensionality of big data sets called principal component analysis, or PCA, involves condensing a great number of variables into a single smaller variable while keeping the majority of the data from the larger set. Accuracy always grieves when an acquired group or set has fewer variables, However, dimensionality reduction seeks to simplify while sacrificing some accuracy. Smaller data sets simplify and accelerate the analysis process for machine learning or prediction systems. without unnecessary variables because they are easy to inspect and visualise. PCA's major objective is to maintain as much information as feasible while minimising the number of variables used to obtain data. Initially, all the variables are standardised to transform into the same scale. Next, the covariance matrix is calculated to determine the variables' correlation. If the correlation is positive, both variables are positively correlated, so both can increase or decrease together. But if the correlation is negative, both variables are inversely correlated. In the next step, find the covariance matrix's eigenvalues and eigenvectors are determined. The data is then recast along the axes of the principal components after determining which main components to keep using a feature vector.

3.2.2. Linear Discriminant Analysis

Supervised classification uses linear discriminant analysis, sometimes referred to as normal discriminant analysis or discriminant function analysis. It as a dimension-reduction technique. It simulates group distinctions like dividing two or more classes. In order to project qualities from higher dimensions onto lower ones, this technique is utilised. For example, we have two classes that must be efficiently separated. Classes can have numerous characteristics, and using only one trait to categorise them may result in some overlapping. So, for correct classification, we need to increase the number of attributes. The creation of a novel axis is based on two ideas. (i) Increase the gap between the mean of the two classes. (ii) Reduce variation within each class. Linear Discriminant Analysis (LDA) is unable to find a new axis that makes both classes linearly separable when the distributions' common mean exists. Non-linear discriminant analysis is utilised in these situations.

This reviewed literature of research papers and different methodologies available for feature selection and extraction techniques provided the basic background for developing a model to categorize the nature of cardio vascular disease more efficiently. The significant proposed model and algorithm are given in Fig. 3 and 4 using different feature selection and extraction techniques.

4. Proposed Work

In proposed model incorporates many phases. In proposed model incorporates many phases. Initially, we collected the heart disease Cleveland dataset, as shown in Table 1, which consists of 14 attributes and 303 individual heart disease records. This dataset contains two classes. If the disease is present, it is labelled with 1 and if the disease is absent, it is labelled with 0. To pre-process the data, we have normalised the data, and missing values are replaced with 0. As this dataset is highly imbalanced, it may hamper the precision of predictive learning algorithms and lead to ignoring the minority class. So, to avoid this problem, resample the dataset randomly. There are two approaches to resample the dataset. Oversampling, or increasing the sample size in the minority class, is one of the approaches. The second is known as under sampling which involves reducing the sample size in the majority class. After this phase, the next is to use feature selection or feature extraction techniques. The feature selection process will be helpful in finding the most advantageous combination of the existing characteristics automatically. Similarly, Feature extraction is a dimension-reduction process that will lessen the number of attributes by creating a new extended attribute compared to the selected one. In both cases, we can improve the accuracy, speed up the training process and reduce the overfitting risk. We have used filter and wrapper methods in this proposed work as feature selection. PCA and LDA are used as feature extraction techniques. After feature selection and feature extraction is performed, a large number of feature subsets are found. However, because all of the feature subsets are irrelevant, the final feature subset is chosen based on performance. Further, we have used different traditional classifiers like Logistic Regression, Naive Bayes, Random Forest, Extreme gradient boost, K-Nearest Neighbour, Decision Tree, and Support vector classifier on a subset of chosen features to categorise the existence or absence of cardiac disease in that particular subgroup. Finally, the performance matrix in terms of accuracy, precision, recall, and F1 score is calculated for each individual classifier. Fig. 4 outlines the suggested model's elaborate construction.

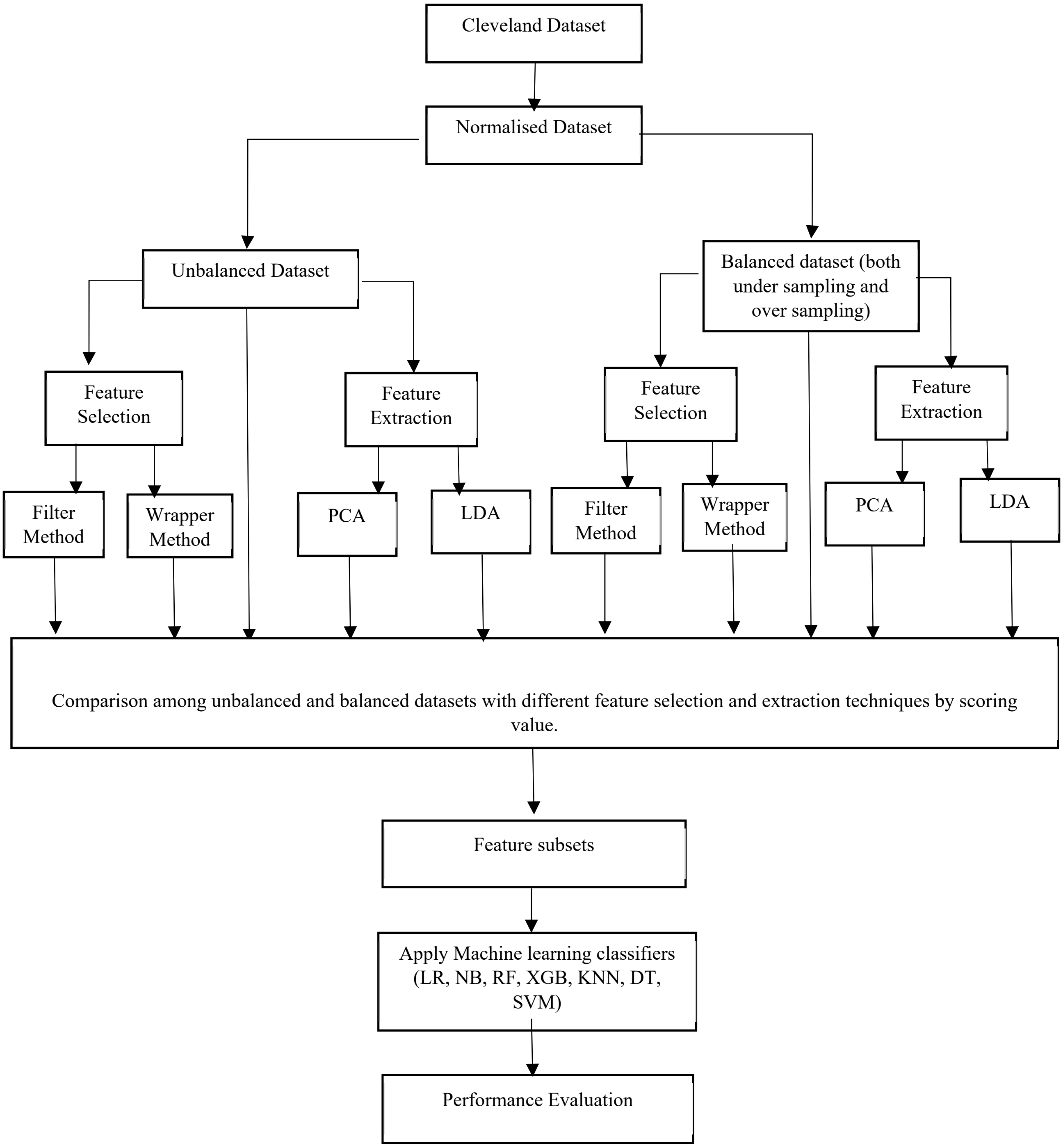

Fig. 3. A detailed architecture of the proposed model

```
1. Input: Original Dataset with n number of features.
2. Pre-processed and normalised the dataset.
3. Balance the dataset using over sampling and under sampling methods.
4. For each n number of features in both balanced and unbalanced datasets (only pre-
   processed and normalised), do
           While (each filter and wrapper method not evaluated)
                   Apply feature selection techniques to each feature column evaluation.
                   Calculate each feature's weight or score.
                   Use machine learning classifiers to assess each feature selection
                   technique on the selected features after choosing the top features
                   based on weight/score.
           Calculate the F1 score, Accuracy, Precision, and Recall.
           End While
   End For
5. For each n number of features in balanced and unbalanced datasets use feature
   extraction PCA and LDA
           Apply machine learning classifiers on selected features and calculate
           Accuracy, Precision, Recall and F1 score
   End For
6. Select the Top three methods based on accuracy and return significant feature
   subsets in both balanced and unbalanced (only pre-processed and normalised).
7. Output: Significant feature subsets and their performances in terms of accuracy,
   precision, recall and F1 score.
```

Fig. 4. The proposed algorithm

Attribute No.	Attribute	Description	Attribute Specification
0	Age	age (in years)	
1	Sex	Sex	1=male,0=female
2	Cp	Subsections of the chest discomfort	Value 1: typical angina Value 2: atypical angina Value 3: non-anginal pain Value 4: asymptomatic
3	Trestbps	Blood pressure at rest (in mmHg)	
4	Chol	Cholestoral including HDL & LDL (in mg/dl)	
5	Fbs	fasting blood sugar >120 mg/dl	1=true; 0=false
6	Restecg	Resting electrocardiographic results	0=normal 1=having ST-T wave abnormality 2=left ventricular hypertrophy
7	Thalach	attainment of the highest pulse rate	
8	Exang	Exercise induced angina	
9	Oldpeak	ST depression induced by exercise relative to rest	
10	Slope	The slope of the peak exercise ST segment	1=upsloping, 2=flat, 3=downsloping
11	Ca	Number of major vessels	
12	Thal	Flaw types	3= normal
13	Num	Final diagnosis	Value 0: <50% diameter narrowing (Normal) Value 1: >50% diameter narrowing (Patient)

Table 1. Data set Description

5. Result and Discussion

Utilizing the Cleveland Clinic Heart Disease Dataset, which is available on the site Cleveland Clinic Heart Disease Dataset | Kaggle, we assessed the performance of the suggested model. Further various experimental research are executed in the Google Colab environment. In this regard, the splitting ratio of the training and testing set is set to be 80:20. Different filter methods like dropping constant, correlation, mutual information classification, mutual information regression, chi-square, fisher score, variance threshold, mean absolute difference, and dispersion ratio are used to perform feature selection. For feature selection, wrapper techniques like the forward, backward, bidirectional, and exhaustive types are employed. Principal component analysis (PCA) and linear discriminant analysis (LDA) were utilised for feature extraction. Further the exhaustive analysis was performed on seven machine learning classifiers Logistic Regression (LR), Naïve Bayes (NB), Random Forest (RF), Extreme Gradient Boost (XGB), K-Nearest Neighbour (KNN), Decision Tree (DT) and Support Vector Classifier (SVC).

In this research, we have considered balanced (both over-sampling and under-sampling) and unbalanced datasets for the experiment. We found that most of the classifiers' performance is better in the under-sampling dataset. So in this section, we have focused only on the results related to the under-sampling dataset. Table 2 depicts the resultant performance for the filter methods considering the under-sampling dataset. The evidence indicates that the SVM shows a high accuracy of 88.82 with the fisher score feature selection method. Similarly the respective performance of feature extraction techniques PCA, LDA and wrapper methods are being represented in table 3 and 4. In table 3, the k-nearest neighbor with LDA feature extraction shows an accuracy of 93.44.

Filter Methods	**Performance Evaluation Metrics**	**LR**	**NB**	**RF**	**XGB**	**KNN**	**DT**	**SVC**
Dropping Constant	Accuracy	85.25	83.61	85.25	85.25	85.25	73.77	85.25
	Precision	0.85 0.86	0.84 0.83	0.82 0.88	0.85 0.86	0.82 0.88	0.68 0.80	0.80 0.90
	Recall	0.81 0.88	0.78 0.88	0.85 0.85	0.81 0.88	0.85 0.85	0.78 0.71	0.89 0.82
	F1 Scores	0.83 0.87	0.81 0.86	0.84 0.87	0.83 0.87	0.84 0.87	0.72 0.75	0.84 0.86
Correlation	Accuracy	83.61	81.97	86.86	83.61	86.89	77.05	83.61
	Precision	0.84 0.83	0.81 0.83	0.83 0.91	0.84 0.83	0.85 0.88	0.71 0.83	0.79 0.88
	Recall	0.78 0.88	0.78 0.85	0.89 0.85	0.78 0.88	0.85 0.88	0.81 0.74	0.85 0.82
	F1 Scores	0.81 0.86	0.79 0.84	0.86 0.88	0.81 0.86	0.85 0.88	0.76 0.78	0.82 0.85
Mutual Information Classification	Accuracy	85.25	85.25	88.52	83.61	88.52	73.77	81.97
	Precision	0.85 0.86	0.82 0.88	0.88 0.89	0.87 0.82	0.86 0.91	0.67 0.82	0.81 0.83
	Recall	0.81 0.88	0.85 0.85	0.85 0.91	0.74 0.91	0.89 0.88	0.81 0.68	0.78 0.85
	F1 Scores	0.83 0.87	0.84 0.87	0.87 0.90	0.80 0.86	0.87 0.90	0.73 0.74	0.79 0.84
Mutual Information Regression	Accuracy	85.25	83.61	85.25	88.52	85.25	78.69	85.25
	Precision	0.85 0.86	0.84 0.83	0.85 0.86	0.88 0.89	0.85 0.86	0.73 0.84	0.85 0.86
	Recall	0.81 0.88	0.78 0.88	0.81 0.88	0.85 0.91	0.81 0.88	0.81 0.76	0.81 0.88
	F1 Scores	0.83 0.87	0.81 0.86	0.83 0.87	0.87 0.90	0.83 0.87	0.77 0.80	0.83 0.87
Chi-Square	Accuracy	70.33	78.69	77.04	81.97	77.05	73.77	72.13
	Precision	0.78 0.82	0.77 0.80	0.74 0.79	0.83 0.81	0.71 0.83	0.72 0.75	0.67 0.77
	Recall	0.78 0.82	0.74 0.82	0.74 0.79	0.74 0.88	0.81 0.74	0.67 0.79	0.74 0.71
	F1 Scores	0.78 0.82	0.75 0.81	0.74 0.79	0.78 0.85	0.76 0.78	0.69 0.77	0.70 0.74
Fisher Score	Accuracy	86.89	85.25	86.89	85.25	86.89	73.77	**88.82**
	Precision	0.88 0.86	0.88 0.84	0.85 0.88	0.85 0.86	0.85 0.88	0.69 0.78	0.86 0.91
	Recall	0.81 0.91	0.78 0.91	0.85 0.88	0.81 0.88	0.85 0.88	0.74 0.74	0.89 0.88
	F1 Scores	0.85 0.89	0.82 0.87	0.85 0.88	0.83 0.87	0.85 0.88	0.71 0.76	0.87 0.90
Variance Threshold	Accuracy	86.89	85.25	83.61	85.25	85.25	78.69	88.52
	Precision	0.88	0.88	0.79	0.85	0.82	0.75	0.86

		0.86	0.84	0.88	0.86	0.88	0.82	0.91
	Recall	0.81 0.91	0.78 0.91	0.85 0.82	0.81 0.88	0.88 0.85	0.78 0.79	0.89 0.88
	F1 Scores	0.85 0.89	0.82 0.87	0.82 0.85	0.83 0.87	0.84 0.87	0.76 0.81	0.87 0.90
Mean Absolute Difference	Accuracy	86.89	85.25	83.61	85.25	85.25	77.05	88.52
	Precision	0.88 0.86	0.88 0.84	0.79 0.88	0.85 0.86	0.82 0.88	0.71 0.83	0.86 0.91
	Recall	0.81 0.91	0.78 0.91	0.85 0.82	0.81 0.88	0.85 0.85	0.81 0.74	0.89 0.88
	F1 Scores	0.85 0.89	0.82 0.87	0.82 0.85	0.83 0.87	0.84 0.87	0.76 0.78	0.87 0.90
Dispersion Ratio	Accuracy	86.89	85.25	83.61	85.25	85.25	77.05	88.52
	Precision	0.88 0.86	0.88 0.84	0.79 0.88	0.85 0.86	0.82 0.88	0.71 0.83	0.86 0.91
	Recall	0.81 0.91	0.78 0.91	0.85 0.82	0.81 0.88	0.85 0.85	0.81 0.74	0.89 0.88
	F1 Scores	0.85 0.89	0.82 0.87	0.82 0.85	0.83 0.87	0.84 0.87	0.76 0.78	0.87 0.90

Table 2. Performance of Feature Selection (Filter methods) in Under-Sampling dataset

Feature Extraction Techniques	**Performance Evaluation Metrics**	**LR**	**NB**	**RF**	**XGB**	**KNN**	**DT**	**SVM**
LDA	Accuracy	91.8	91.8	88.52	91.8	**93.44**	86.88	91.8
	Precision	0.89 0.94	0.89 0,94	0.88 0.89	0.96 0.89	0.93 0.94	0.85 0.88	0.89 0.94
	Recall	0.93 0.91	0.93 0.91	0.85 0.91	0.85 0.97	0.93 0.94	0.85 0.88	0.93 0.91
	F1 Scores	0.91 0.93	0.91 0.93	0.87 0.90	0.90 0.93	0.93 0.94	0.85 0.88	0.91 0.93
PCA	Accuracy	80.33	78.69	80.33	**85.25**	78.69	78.69	80.33
	Precision	0.78 0.82	0.77 0.80	0.76 0.84	0.88 0.84	0.73 0.84	0.77 0.80	0.80 0.81
	Recall	0.78 0.82	0.74 0.82	0.81 0.79	0.78 0.91	0.81 0.76	0.74 0.82	0.74 0.85
	F1 Scores	0.78 0.82	0.75 0.81	0.79 0.82	0.82 0.87	0.77 0.80	0.75 0.81	0.77 0.83

Table 3. Performance of Feature Extraction (LDA and PCA) in Under Sampling

Wrapper Methods	**Performance Evaluation Metrics**	**LR**	**NB**	**RF**	**XGB**	**KNN**	**DT**	**SVC**
Forward	Accuracy	83.78	84.68	**95.95**	90.09	92.34	**98.2**	**95.5**
	Precision	0.85	0.84	0.96	0.86	1.00	0.99	0.95
	Recall	0.90	0.89	0.97	0.92	0.84	0.99	0.96
	F1 Score	0.84	0.84	0.96	0.89	0.92	0.98	0.96
Backword	Accuracy	85.58	84.68	**96.4**	88.74	93.24	**98.2**	**95.95**
	Precision	0.85	0.83	0.95	0.86	1.00	0.99	0.95
	Recall	0.90	0.90	0.99	0.93	0.85	1.00	0.96
	F1 Score	0.85	0.85	0.96	0.89	0.93	0.98	0.96
Bidirectional	Accuracy	85.58	85.58	**96.4**	89.19	93.69	**98.2**	**95.95**
	Precision	0.85	0.85	0.96	0.88	1.00	0.99	0.95
	Recall	0.90	0.90	0.99	0.94	0.87	1.00	0.97
	F1 Score	0.86	0.85	0.96	0.89	0.93	0.98	0.96
Exhaustive	Accuracy	78.69	78.69	83.61	85.25	80.33	73.77	77.05
	Precision	0.84	0.82	0.85	0.90	0.87	0.82	0.81
	Recall	0.7676	0.79	0.85	0.82	0.76	0.68	0.76
	F1 Score	0.80	0.81	0.85	0.86	0.81	0.74	0.79

Table 4. Performance of Wrapper Methods in the under-sampling dataset

As shown in Table 4, the accuracy of decision tree (DT), random forest (RF) and support vector classifier (SVC) are the top three classifiers that show maximum accuracy among others. DT with forward, backward and bidirectional feature selections are the highest accuracy 98.2. In a significant comparison, the accuracy of our proposed model is found to be better than the recent studies [33] and [34], which were 88.52 and 85.8, respectively.

A detailed performance of machine learning classifiers (LR, NB, RF, XGB, KNN, DT and SVC) with wrapper methods is shown in the Fig. 5. According to Fig. 6, the most significant features subset using forward selection and DT is (0, 1, 2, 3, 4, 6, 7, 10, 11, 12) that represents the features “age”, “sex”, “cp”, “trestbps”, “chol”, “restecg”, “thalach”, “slop”, “ca”, “thal” for predicting cardiovascular disease which is showing the accuracy of 98.2. Similarly, DT with bidirectional features selection subset is (0, 1, 2, 3, 4, 7, 8, 9, 11) i.e. “age”, “sex”, “cp”, “trestbps”, “chol”, “thalach”, “exang”, “oldpeak”, “ca” which is showing same accuracy 98.2 as depicted in Fig. 7. Fig. 8 demonstrates the relevant features subset considering backward feature selection and DT is again (0, 1, 2, 3, 4, 7, 8, 9, 11) i.e. “age”, “sex”, “cp”, “trestbps”, “chol”, “thalach”, “exang”, “oldpeak”, “ca”. Similarly, the significant feature subset in the case of RF and SVC using different wrapper methods are shown in Fig. 9 to 14.

As we conclude feature selection using forward, backward, and bidirectional, showing the highest accuracy, further different error calculation functions have been taken into account. We have considered five loss functions: mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), mean squared logarithmic error (MSLE) and log loss error (LLE). The anticipated lesser number for error demonstrates the model's improved correctness. Table 5 includes a detailed analysis of the wrapper feature selection with different classifiers.

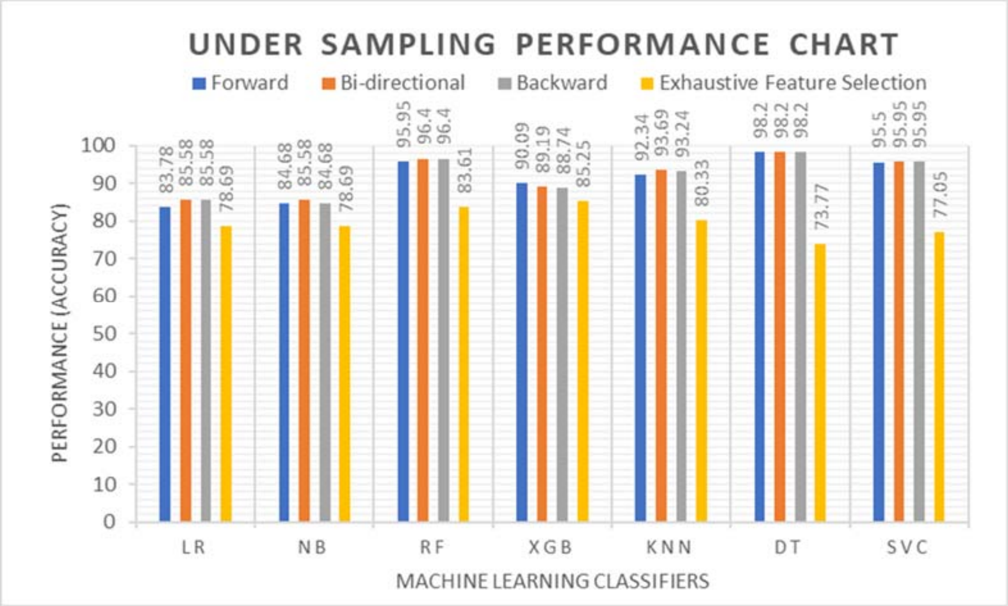

Fig. 5. Performance of machine learning classifiers in wrapper methods in the under-sampling in terms of accuracy.

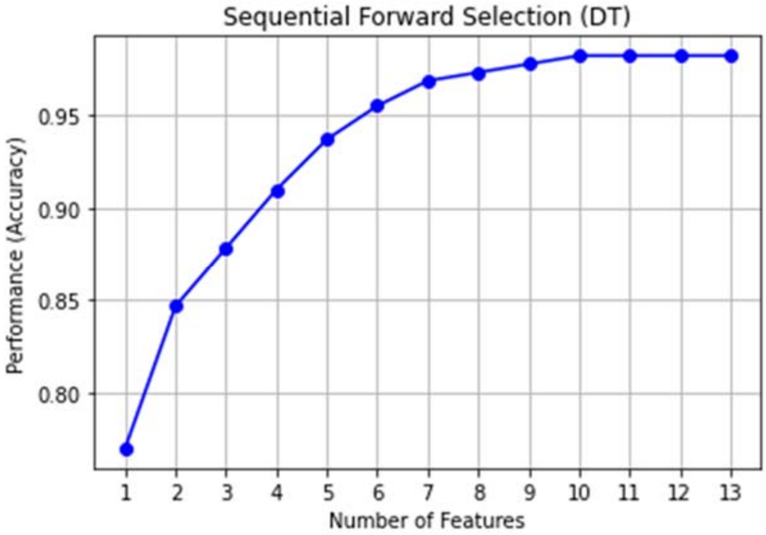

Fig. 6. Performance of feature subset (0, 1, 2, 3, 4, 6, 7, 10, 11, 12) i.e "age", "sex", "cp", "trestbps", "chol", "restecg", "thalach", "slop", "ca", "thal" for DT using forward selection method.

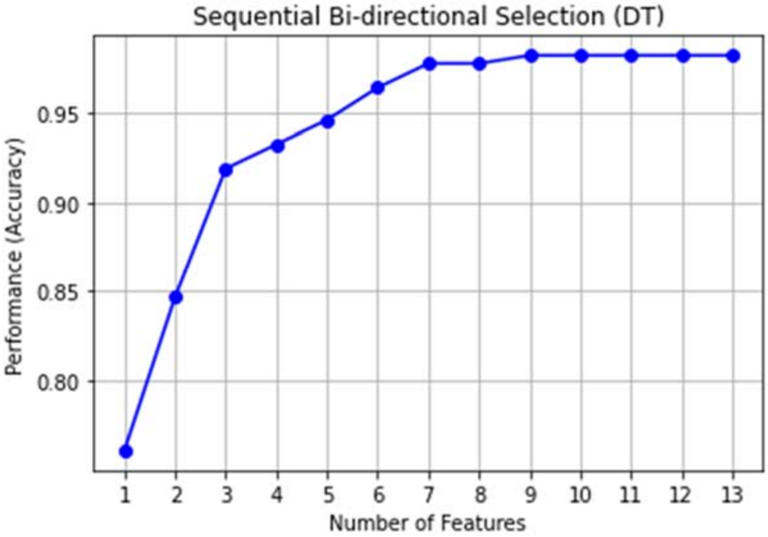

Fig. 7. The performance of features subset (0, 1, 2, 3, 4, 7, 8, 9, 11) i.e. "age", "sex", "cp", "trestbps", "chol", "thalach", "exang", "oldpeak", "ca" for DT using bidirectional selection method.

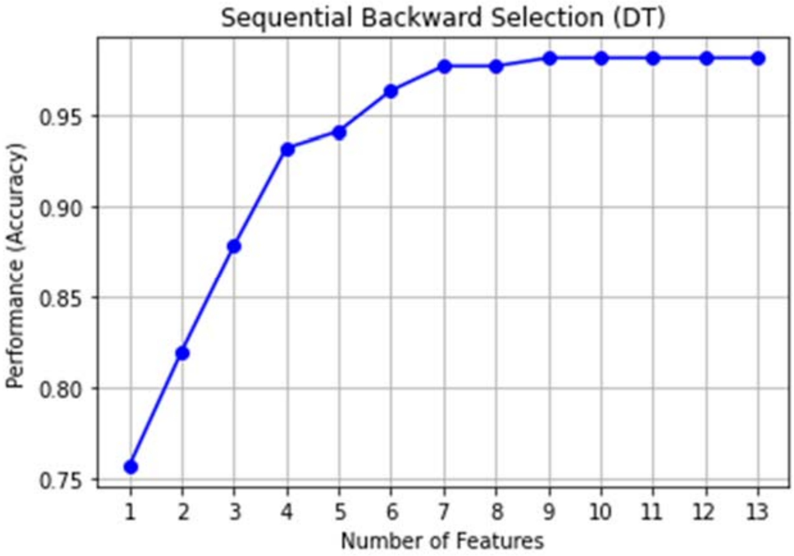

Fig. 8. The performance of the feature subset (0, 1, 2, 3, 4, 7, 8, 9, 11) i.e "age", "sex", "cp", "trestbps", "chol", "thalach", "exang", "oldpeak", "ca" for DT using backword selection method.

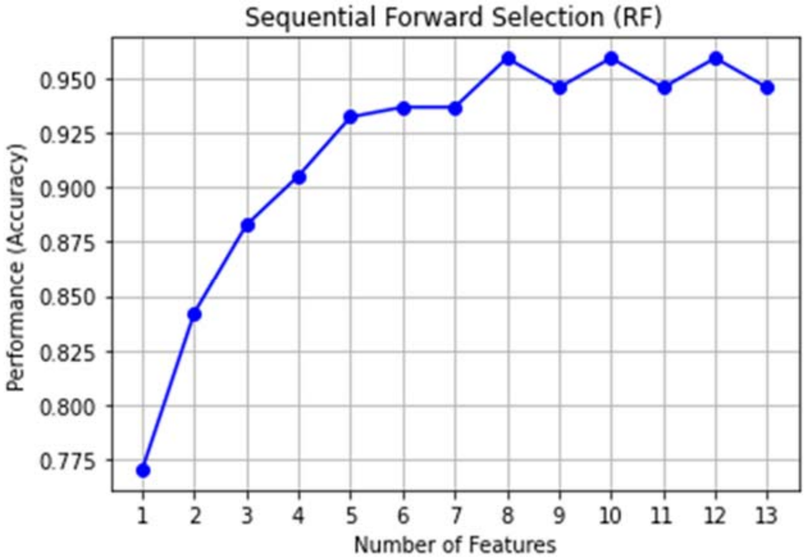

Fig. 9. The performance of features subset (0, 2, 3, 5, 7, 10, 11, 12) i.e. "age", "cp", "trestbps", "fbs", "thalach", "slope", "ca", "thal" for RF using forward selection method.

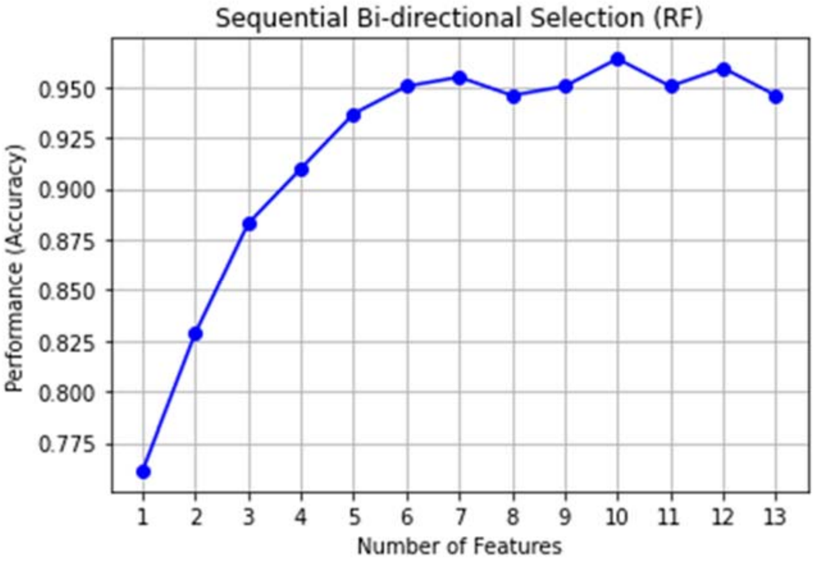

Fig. 10. The performance of features subset (0, 2, 3, 4, 5, 6, 7, 9, 10, 11) i.e. “age”, “cp”, “trestbps”, “chol”, “fbs”, “restecg”, “thalach”, “oldpeak”, “slope”, “ca” for RF using bidirectional selection method.

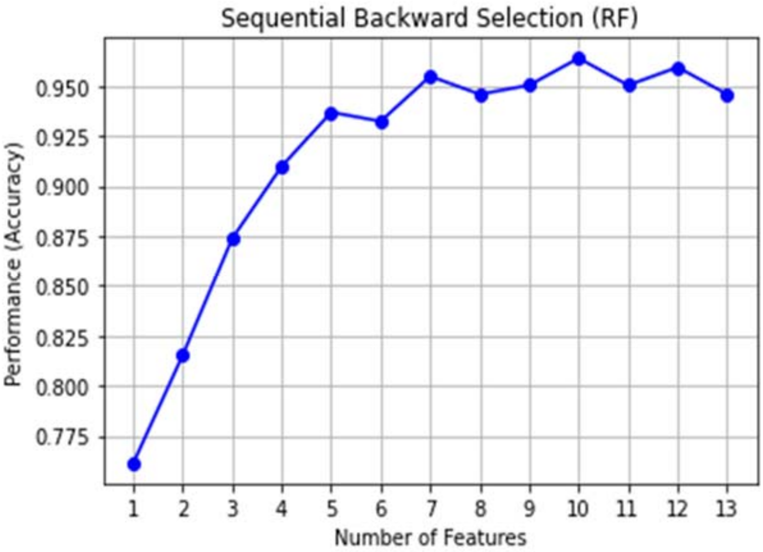

Fig. 11. The performance of features subset (0, 2, 3, 4, 5, 6, 7, 9, 10, 11) i.e. “age”, “cp”, “trestbps”, “chol”, “fbs”, “restecg”, “thalach”, “oldpeak”, “slope”, “ca” for RF using backward selection method.

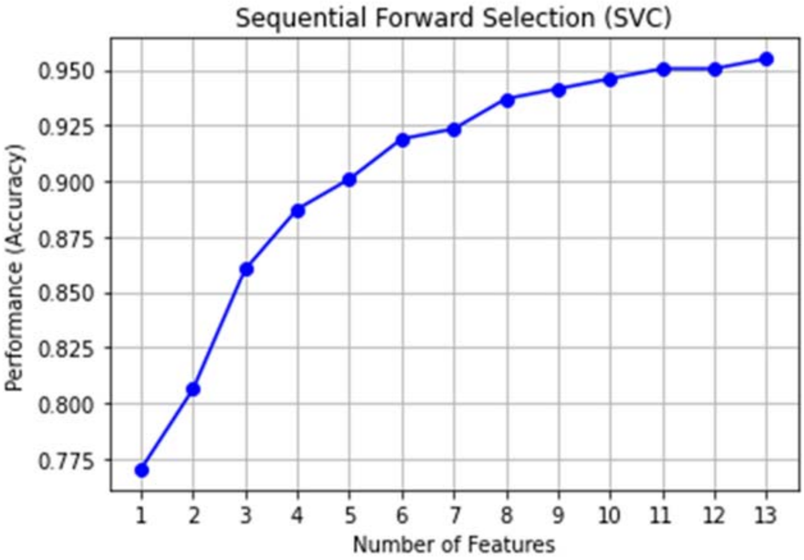

Fig. 12. The performance of features subset (0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12) i.e "age", "sex", "cp", "trestbps", "chol", "fbs", "restecg", "thalach", "exang", "oldpeak", "slope", "ca", "thal" for SVC using forward selection method.

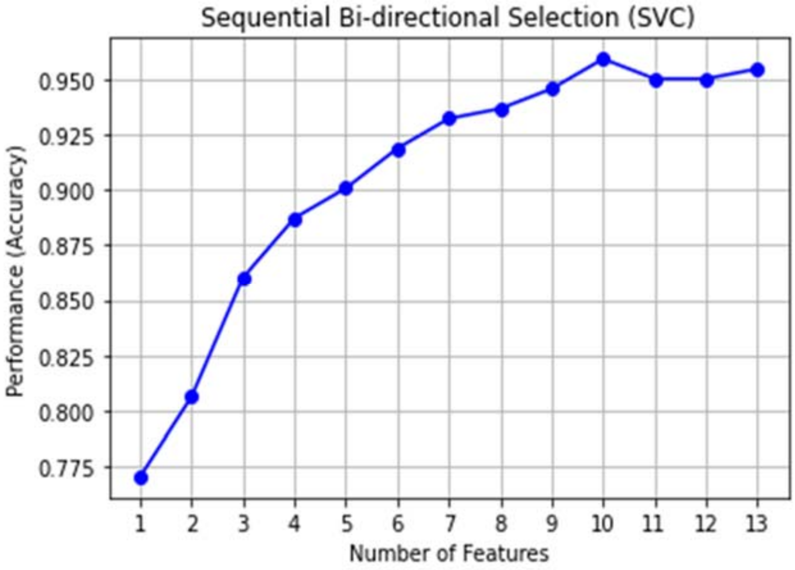

Fig. 13. The performance of features subset (0, 1, 2, 3, 4, 5, 8, 10, 11, 12) i.e. "age", "sex", "cp", "trestbps", "chol", "fbs", "exang", "slope", "ca", "thal" for SVC using bidirectional selection method.

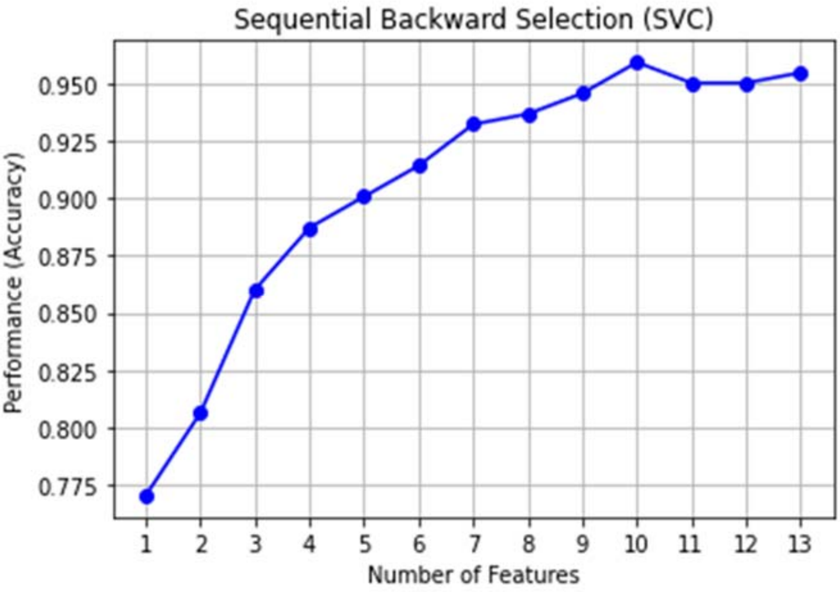

Figure 14. The performance of features subset (0, 1, 2, 3, 4, 5, 8, 10, 11, 12) i.e. “age”, “sex”, “cp”, “trestbps”, “chol”, “fbs”, “exang”, “slope”, “ca”, “thal” for SVC using backword selection method

Loss Function	**Feature selection methods**	**LR**	**NB**	**RF**	**XGB**	**KNN**	**DT**	**SVC**
MAE	Forward	0.162162	0.153153	0.040541	0.099099	0.076577	0.018018	0.045045
	Bidirectional	0.144144	0.144144	0.036036	0.108108	0.063063	0.018018	0.040541
	Backword	0.144144	0.153153	0.036036	0.112613	0.067568	0.018018	0.040541
	Exhaustive	0.213115	0.213115	0.163934	0.147541	0.196721	0.262295	0.229508
MSE	Forward	0.162162	0.153153	0.040541	0.099099	0.076577	0.018018	0.045045
	Bidirectional	0.144144	0.144144	0.036036	0.108108	0.063063	0.018018	0.040541
	Backword	0.144144	0.153153	0.036036	0.112613	0.067568	0.018018	0.040541
	Exhaustive	0.213115	0.213115	0.163934	0.147541	0.196721	0.262295	0.229508
RMSE	Forward	0.402694	0.391348	0.201347	0.3148	0.276725	0.134231	0.212238
	Bidirectional	0.379663	0.379663	0.189832	0.328798	0.251124	0.134231	0.201347
	Backword	0.379663	0.391348	0.189832	0.335578	0.259938	0.134231	0.201347
	Exhaustive	0.461644	0.461644	0.404888	0.384111	0.443533	0.512148	0.47907
MSLE	Forward	0.077911	0.073583	0.019478	0.054105	0.036791	0.008657	0.019478
	Bidirectional	0.069254	0.069254	0.017314	0.051941	0.030299	0.008657	0.019478
	Backword	0.069254	0.073583	0.017314	0.054105	0.032463	0.008657	0.019478
	Exhaustive	0.102392	0.102392	0.078763	0.070887	0.094515	0.12602	0.110268
LLE	Forward	0.342498	0.431593	0.207223	0.626798	0.121769	0.041288	0.201427
	Bidirectional	0.342498	0.419381	0.202916	0.626798	0.112402	0.029692	0.201559
	Backword	0.342498	0.419381	0.202916	0.626798	0.112402	0.029692	0.19362
	Exhaustive	7.360788	7.360801	5.66216	5.095924	6.794566	9.059417	7.927011

Table 5. Error Calculation of different classifiers using MAE, MSE, RMSE, MSLE and LLE

6. Conclusion

In this research work, we contemplated at how feature extraction and selection affected the precision of cardiovascular disease. The Cleveland heart disease dataset from Kaggle is being used as the beam in the analysis.

Experiments were performed on balanced and unbalanced datasets with feature selection (filter and wrapper methods) and feature extraction techniques. In filter methods dropping constant, correlation, mutual information classification, mutual information regression, chi-square, fisher score, variance threshold, mean absolute difference, and dispersion ratio are used. Similarly, in the case of wrapper methods, forward, backward, bidirectional, and exhaustive methods are employed for the attribute selection. Correspondingly, principal component analysis (PCA) and linear discriminant analysis (LDA) have been utilised for attribute extraction. The analysis was performed on seven machine learning classifiers. The decision tree (DT) shows a higher accuracy 98.2 in the case of forward, backward, and bidirectional feature selection techniques among other classifiers, and finally compared with other recent works.

References

[1] Felman, A. (2018): Everything you need to know about heart disease. Medical News Today, https://www.medicalnewstoday.com/articles/237191.

[2] Thomas, J.; Princy, R. (2016): Human heart disease prediction system using data mining techniques. 2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT), pp. 1-5.

[3] Saboor, A.; Usman, M.; Ali, S.; Samad, A.; Abrar, M. F.; Ullah N. (2022): A Method for Improving Prediction of Human Heart Disease Using Machine Learning Algorithms, Mobile Information Systems, vol. 2022.

[4] Desai, R.J.; Wang, S.V.; Vaduganathan, M.; Evers, T.; Schneeweiss, S. (2020): Comparison of Machine Learning Methods With Traditional Models for Use of Administrative Claims With Electronic Medical Records to Predict Heart Failure Outcomes. JAMA Netw Open. 2020;3(1):e1918962. doi:10.1001/jamanetworkopen.2019.18962.

[5] Alaa, AM.; Bolton, T.; Di, Angelantonio, E.; Rudd, JHF.; van der Schaar, M. (2019): Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. PLoS One. 14(5):e0213653. doi: 10.1371/journal.pone.0213653. PMID: 31091238; PMCID: PMC6519796.

[6] Hussain, W.; Aldahiri, A.; Alrashed, B. (2021): Trends in Using IoT with Machine Learning in Health Prediction System. Forecasting 2021, 3, 181–207.

[7] Kishor, A.; Chakraborty, C. (2021): Artificial Intelligence and Internet of Things Based Healthcare 4.0 Monitoring System. Wirel. Pers. Commun., 2, 1–17.

[8] Ristevski, B.; Snezana, S. Healthcare and Medical Big Data Analytics; (2021) Elsevier: Amsterdam, The Netherlands.

[9] Lee, S.S.; Ae, KK; Kim, D.; Lim, Y.; Yang, P.; Yi, J.; Kim, M.; Kwon, K.; Pyun, W.B.; Joung, B.; et al. (2017): Clinical Implication of an Impaired Fasting Glucose and Prehypertension Related to New Onset Atrial Fibrillation in a Healthy Asian Population without Underlying Disease. Eur. Heart J. 2017, 38, 2599–2607.

[10] Alsunaidi, S.J.; Almuhaideb, A.M.; Ibrahim, N.M.; Shaikh, F.S.; Alqudaihi, K.S.; Alhaidari, F.A.; Khan, I.U.; Aslam, N.; Alshahrani, M.S. (2021): Applications of Big Data Analytics to Control COVID-19 Pandemic. Sensors 2021, 21, 2282.

[11] Ali, M.M.; Paul, B.K.; Ahmed, K.; Bui, F.M.; Quinn, J.M.; Moni, MA. (2021): Heart Disease Prediction Using Supervised Machine Learning Algorithms: Performance Analysis and Comparison. Comput. Biol. Med. 2021, 136, 104672.

[12] Dogan, O.; Tiwari, S.; Jabbar, M.A.; Guggari, S. (2021): A Systematic Review on AI/ML Approaches against COVID-19 Outbreak. Complex. Intell. Syst. 2021, 7, 2655–2678.

[13] Aggrawal, R.; Pal, S. (2020), "Sequential feature selection and machine learning algorithm-based patient's death events prediction and diagnosis in heart disease," SN Computer Science, vol. 1, no. 6.

[14] Gao, XY.; Ali, A.A.; Hassan, S.H.; Anwar,E.M., (2021): Improving the accuracy for analysing heart diseases prediction based on the ensemble method, Complexity, vol. 2021, Article ID 6663455, 10 pages.

[15] H. Takci, (2018) Improvement of heart attack prediction by the feature selection methods," Turkish Journal of Electrical Engineering and Computer Sciences, vol. 26, pp. 1–10.

[16] Latha, C. B. C. and Jeeva, S. C. (2019): Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques, Informatics in Medicine Unlocked, vol. 16.

[17] Garate-Escamila, A. K.,. Hassani, A. E., and Andres, E., (2020): Classification models for heart disease prediction using feature selection and PCA, Informatics in Medicine Unlocked, vol. 19.

[18] Spencer,R., Thabtah, F., Abdelhamid, N., Thompson, M. (2020): Exploring feature selection and classification methods for predicting heart disease. DIGITAL HEALTH. 2020;6.

[19] Senan, E. M., Al-Adhaileh, M. H., Alsaade, F. W. et al. (2021): Diagnosis of chronic kidney disease using effective classification algorithms and recursive feature elimination techniques, Journal of Healthcare Engineering, vol. 2021, Article ID 1004767, 10 pages, 2021.

[20] Almansour, N. A., Syed, H. F., Khayat N. R. et al. (2019): Neural network and support vector machine for the prediction of chronic kidney disease: a comparative study, Computers in Biology and Medicine, vol. 109, pp. 101–111.

[21] Schiezaro, M., Pedrini, H. (2013): Data feature selection based on Artificial Bee Colony algorithm. EURASIP J. Image Video Process. 2013(1), 47 (2013). https://doi.org/10.1186/1687-5281-2013-47 35.

[22] Chandrashekar, G., Sahin, F. (2013): A survey on feature selection methods. Comput. Electr. Eng. 40(1), 16–28 (2014). https://doi.org/10. 1016/j.compeleceng.2013.11.024 36.

[23] Liu, H., Yu, L. (2005): Toward integrating feature selection algorithms for classification and clustering. IEEE Trans. Knowl. Data Eng. 17(4), PP. 491–502.

[24] Guyon, I., Elisseeff, A. (2003): An introduction to variable and feature selection. J. Mach. Learn. Res. 3, 1157–1182.

[25] Tuv, E., Borisov, A., Runger, G., Torkkola, K. (2009): Feature selection with ensembles, artificial variables, and redundancy elimination. J. Mach. Learn. Res. 10, 1341–1366.

[26] Vergara, J.R., Estévez, P.A. (2014): A review of feature selection methods based on mutual information. Neural Comput. Appl. 24(1), 175–186.

[27] Jovi'c, A., Brki'c, K., Bogunovi'c, N. (2015): A review of feature selection methods with applications. In: 2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO) (pp. 1200–1205), IEEE.

[28] Bolón-Canedo, V., Sánchez-Maroño, N., Alonso-Betanzos, A. (2013): A review of feature selection methods on synthetic data. Knowl. Inf. Syst. 34(3), 483–519.

[29] Remeseiro, B., Bolon-Canedo, V. (2019): A review of feature selection methods in medical applications. Comput. Biol. Med. 112, 103375 (2019). https://doi.org/10.1016/j.compbiomed.2019.103375.

[30] Shah, F.P., Patel, V. (2016): A review on feature selection and feature extraction for text classification. In: 2016 international conference on wireless communications, signal processing and networking (WiSPNET) (pp. 2264–2268), IEEE.
[31] Mwadulo, M.W. (2016): A review on feature selection methods for classification tasks, International Journal of Computer Applications Technology and Research 5(6):395-402.
[32] Nouri-Moghaddam, B., Ghazanfari, M., Fathian, M. (2021): A novel bioinspired hybrid multi-filter wrapper gene selection method with ensemble classifier for microarray data. Neural Comput. Appl.
[33] Dissanayake, K., Gapar Md, Johar Md.(2021): Comparative Study on Heart Disease Prediction Using Feature Selection Techniques on Classification Algorithms, *Applied Computational Intelligence and Soft Computing*, vol. 2021, 17 pages.
[34] Reddy, K. V. V. , Elamvazuthi, I., Aziz, A. A., Paramasivam, S., Chua, H. N. (2020): Heart Disease Risk Prediction using Machine Learning with Principal Component Analysis, 2020 8th International Conference on Intelligent and Advanced Systems (ICIAS*)*,pp.1-6.

Authors Profile

Subhalaxmi Das is a Ph.D. Research Scholar in the Department of Computer Science and Applications, Utkal University, Bhubaneswar, Odisha, India. She received B.Tech in Computer Science and Engineering from Biju Patnaik University of Technology, Odisha, India and M.Tech in Computer Science & Engineering from KIIT University, India. She has 12 years of teaching experience. She has published some research papers in International Journals and Conferences. Her special fields of interest included Machine Learning, Soft Computing, Image Processing and Data Mining.

Professor Sateesh Kumar Pradhan completed his Ph.D. in Computer Science from Berhampur University, Berhampur, Odisha, India. He was the Former Professor & Head, P.G. Department of Computer Science, and Former Dean Faculty of Engineering, Utkal University, Bhubaneswar, Odisha, India. Twenty-two research scholars have been awarded Ph.D. degree in the area of E-Commerce, Mobile Ad hoc Networks, Intrusion Detection, Computer Forensic, Web Services, Parallel Task Scheduling, Medical Data Mining, Software Testing, Internet of Things, Brain Image Classification, Optical Character Recognition of Odia Document etc. under his supervision and currently eight scholars are continuing for Ph. D. degree under his supervision and guidance. There are more than forty Journal Publications and more than Sixty Conference Publications to his credit. He also served as Senior Faculty in Computer Engineering, King Khalid University, Abha, Saudi Arabia (2006-2011) and Faculty, Computer Science, Berhampur University, Odisha, India (1987-2000).

Sujogya Mishra received Ph. D. in Computer Science from Utkal University, India. He is currently working as an Assistant Professor in the department of Mathematics and Computing in Odisha University of Technology and Research, Bhubaneswar, Odisha, India. He has published more than 26 publications in reputed international Scopus based high reputed Journals in the areas of Soft Computing, Rough Set Theory, Machine Learning and Bioinformatics which are the subjects of his research interest. He has over 25 years of teaching and research experiences.

Dr. Sipali Pradhan completed her Ph.D. in Computer Science & IT from North Orissa University, Baripada , Odisha, India in the year 2019. The title of her Ph. D. thesis is “Rea-Time HealthCare Using Internet of Things”. Dr. Sipali Pradhan completed her M. Tech in Computer Science & Engineering from Utkal University, Bhubaneswar, India in the year 2015 before joining North Orissa University for Ph.D. programme. She is current working as Assistant Professor, Department of Computer Science, RBVRR Women’s College (Osmania University), Hyderabad, India. There are more than ten Journal & Conference Publications to her credit.

Pradyumna Kumar Pattnaik is currently working as an Assistant Professor in the department of Mathematics and Computing in Odisha University of Technology and Research, Bhubaneswar, Odisha, India. He has more than 46 publications in reputed international Scopus based high reputed Journals. He has over 22 years of teaching and research experiences.