International Journal of Big Data and Analytics in Healthcare
Volume 6 • Issue 1 • January-June 2021

Analysis of Heart Disease Using Parallel and Sequential Ensemble Methods With Feature Selection Techniques:

Heart Disease Prediction

Dhyan Chandra Yadav, Veer Bahadur Singh Purvanchal University, Jaunpur, India

Saurabh Pal, Veer Bahadur Singh Purvanchal University, Jaunpur, India

https://orcid.org/0000-0001-9545-7481

ABSTRACT

This paper has organized a heart disease-related dataset from UCI repository. The organized dataset describes variables correlations with class-level target variables. This experiment has analyzed the variables by different machine learning algorithms. The authors have considered prediction-based previous work and finds some machine learning algorithms did not properly work or do not cover 100% classification accuracy with overfitting, underfitting, noisy data, residual errors on base level decision tree. This research has used Pearson correlation and chi-square features selection-based algorithms for heart disease attributes correlation strength. The main objective of this research to achieved highest classification accuracy with fewer errors. So, the authors have used parallel and sequential ensemble methods to reduce above drawback in prediction. The parallel and serial ensemble methods were organized by J48 algorithm, reduced error pruning, and decision stump algorithm decision tree-based algorithms. This paper has used random forest ensemble method for parallel randomly selection in prediction and various sequential ensemble methods such as AdaBoost, Gradient Boosting, and XGBoost Meta classifiers. In this paper, the experiment divides into two parts: The first part deals with J48, reduced error pruning and decision stump and generated a random forest ensemble method. This parallel ensemble method calculated high classification accuracy 100% with low error. The second part of the experiment deals with J48, reduced error pruning, and decision stump with three sequential ensemble methods, namely AdaBoostM1, XG Boost, and Gradient Boosting. The XG Boost ensemble method calculated better results or high classification accuracy and low error compare to AdaBoostM1 and Gradient Boosting ensemble methods. The XG Boost ensemble method calculated 98.05% classification accuracy, but random forest ensemble method calculated high classification accuracy 100% with low error.

KEYWORDS

AdaBoost, Chi-Square, Decision Stump Algorithm, Gradient Boosting, J48 Algorithm, Pearson Correlation, Random Forest Classifiers, Reduced Error Pruning, XGBoost

DOI: 10.4018/IJBDAH.20210101.oa4

1. INTRODUCTION

Virani SS et al., (2020), introduced about heart organ importance. The heart is an important organ of human body in which various vessels support in blood flows to different parts of the body. Our heart is a complex structure consisting of muscles. It suffers from various types of diseases in the heart by more or less blood secretion. The cardiovascular is a heart diseases and the main cause of cardiovascular diseases is poor lifestyle, stress, coma, not exercising and irregular food intake. These life style leads to disability on heart and invites serious heart related diseases. So, healthy heart is very important for human life because it is necessary to keep the body healthy. If a person has been faced the problem of heart attack, then he should change his lifestyle to keep the heart healthy. Sometimes heart diseases show some of their symptoms then we should not avoid it easily.

1.1. Types of Heart Disease

1.1.1. Coronary Artery Disease

This disease is very common, which is caused due to the accumulation of arteries in the arteries and stops the flow of blood in the heart and brings the heart to a state of zero and increases the risk of stroke.

1.1.2. Hyper Tensile Heart Disease

This disease is caused by high blood flow in the arteries, as a result of which the blood vessels become heavy and the incidence related to heart disease problem.

1.1.3. Rheumatic Heart Disease

This disease arises due to damaged heart valve and the patient faced coma as heart disease which is associated with rheumatic fever. Due to this disease, the tissues connecting the brain, the joints of the brain and the skin get affected and these organs become inflamed.

1.1.4. Congenital Heart Disease

This disease is visible due to the lack of proper heart structure at birth, because it is not able to be in normal state due to bad heart condition.

1.2. The Main Causes of Heart Disease

- Increased cholesterol
- Smoking more
- Excess alcohol consumption
- Maintaining a state of constant tension
- Heredity
- Excess body fat
- High blood pressure
- Stretching with pain due to artery block in the heart and increasing the possibility of heart attack
- Excess of nausea
- Chest irritation and imbalance in digestive functions
- Excess of pain in the hands and down the left shoulder
- Excessive phlegm and white or pink color
- Excess of sweat from the body while lack of physical activity

1.3. Algorithms Description

1.3.1. J48

Gupta A et al., (2020), introduced about decision tree and their performance. C4.5 (J48) is a decision tree algorithm that was developed by Ross Quinlan. C4.5 is an extend algorithm of Quinlan's ID3 and the decision trees generated by C4.5 can be used for prediction in data classification. The main objective of this algorithm to solve both classification and regression problems also lead to overfitting of the data (Table 1).

Table 1. Representation of J48 algorithm experimental performance on heart disease attributes

```
cp <= 0
| ca <= 0
| | thal <= 2
| | | thalach <= 118: zero (8.0)
| | | thalach > 118
| | | | exang <= 0
| | | | | chol <= 315: one (68.0)
| | | | | chol > 315
| | | | | | age <= 61: zero (4.0)
| | | | | | age > 61: one (3.0)
| | | | exang > 0
| | | | | restecg <= 0: one (12.0)
| | | | | restecg > 0
| | | | | | slope <= 1
| | | | | | | trestbps <= 115: one (3.0)
| | | | | | | trestbps > 115: zero (21.0)
| | | | | | slope > 1: one (6.0)
| | | | | | | | | | age <= 65: one (10.0)
| | | | | | | | | | age > 65: zero (4.0)
| | | | | | | | | ca > 0: zero (12.0)
| | | | | | | | ca > 2: one (7.0)
| | | | | | | fbs > 0: one (16.0)
| | | | | | chol > 244: zero (37.0)
| | | age > 68: one (22.0)
| oldpeak > 1.9
| | slope <= 0: one (12.0)
| | slope > 0
| | | ca <= 0
| | | | oldpeak <= 2.6: one (6.0)
| | | | oldpeak > 2.6: zero (10.0)
| | | ca > 0: zero (25.0)
Number of Leaves: 48
Size of the tree: 95
```

1.3.2. Reduced Error Pruning

Catherine OO (2020), introduced about Clinical Diagnostic System Driven by Reduced Error. The main objective of reduced pruning tree to simplifying a decision tree by removing branches of the decision tree that is uncritical and redundant to classify instances. The advantage of reduced pruning tree is to reduce the size of decision trees by removing parts of the tree. It can not only reduce the size of the tree but also improve the classification accuracy of unseen objects (Table 2).

1.3.3. Decision Stump

Palad EB et al., (2020), introduced about decision stump machine learning classifiers in data science. It is used to generate one-level decision tree. The decision stump contains one internal node (the root) which is finally connected to the terminal nodes (its leaves). The decision stump prediction based on the value single input feature and depends upon possible several variations. This algorithm Used components weak learners or base learners in ensemble techniques parallel and sequential (Table 3).

1.3.4. Random Forest Ensemble Method

Kaandorp ML, Dwight RP., (2020), introduced about random forests algorithm work as randomly decision forests in machine learning. We can use random forest as an ensemble method for better

Table 2. Representation of REP algorithm experimental performance on heart disease attributes

```
cp < 0.5
| ca < 0.5
| | exang < 0.5
| | | thal < 2.5
| | | | thalach < 96.5: zero (2/0) [2/0]
| | | | thalach >= 96.5: one (50/1) [25/3]
| | | thal >= 2.5
| | | | restecg < 0.5: zero (11/0) [2/0]
| | | | restecg >= 0.5
| | | | | age < 41.5: zero (4/0) [4/0]
| | | | | age >= 41.5: one (8/0) [1/0]
| | sex >= 0.5
| | | chol < 245.5
| | | | oldpeak < 2.4
| | | | | ca < 0.5
| | | | | | age < 65.5: one (21/0) [11/0]
| | | | | | age >= 65.5
| | | | | | | age < 68.5: zero (2/0) [2/0]
| | | | | | | age >= 68.5: one (3/0) [0/0]
| | | | | ca >= 0.5
| | | | | | thal < 2.5: zero (10/3) [5/0]
| | | | | | thal >= 2.5: one (8/1) [6/3]
| | | | oldpeak >= 2.4: zero (9/0) [5/0]
| | | chol >= 245.5: zero (29/4) [20/2]
Size of the tree: 71
```

Table 3. Representation of DS algorithm experimental performance on heart disease attributes

cp <= 0.5	
zero	one
0.7545271629778671	0.2454728370221328
cp > 0.5	
zero	one
0.23484848484848486	0.7651515151515151
cp is missing	
zero	one
0.4868292682926829	0.5131707317073171

classification, regression as decision tree forest model. We can use random forest as parallel ensemble method known as bagging Meta classifier algorithm. This algorithm tries to create an uncorrelated forest of trees by weak learners to strong learners and provide more accurate than that of any individual tree (Table 4).

Table 4. Representation of RF algorithm experimental performance on heart disease attributes

```
cp < 0.5
| thal < 2.5
| | ca < 0.5
| | | restecg < 0.5
| | | | oldpeak < 0.1
| | | | | trestbps < 134
| | | | | | age < 55.5: one (9/0)
| | | | | | age >= 55.5: zero (4/0)
| | | | | trestbps >= 134: one (12/0)
| | | | oldpeak >= 0.1: one (25/0)
| | | | | | | age >= 46
| | | | | | | | trestbps < 114: zero (4/0)
| | | | | | | | trestbps >= 114: one (25/0)
| | | | | | thal >= 2.5
| | | | | | | trestbps < 171: one (14/0)
| | | | | | | trestbps >= 171: zero (3/0)
| | | | | age >= 56.5
| | | | | | cp < 2.5
| | | | | | | chol < 238.5: zero (4/0)
| | | | | | | chol >= 238.5
| | | | | | | | age < 66: zero (8/0)
| | | | | | | | age >= 66: one (6/0)
| | | | | | cp >= 2.5: zero (3/0)
| | | | oldpeak >= 2.55: zero (4/0)
Size of the tree: 135
```

1.3.5. AdaBoostM1 Ensemble Method

Çınar A et al., (2020), introduced about adaptive boosting is an iterative ensemble Meta classifiers algorithm. This method has organized by Yoav Freund and Robert Schapire in 1996. The main objective of this algorithm to achieved high accuracy or increase retrained the weak learners for particular data into strong learners (Table 5).

1.3.6. Gradient Boosting Ensemble Method

Ma B. et al., (2020), introduced about Meta classifier ensemble method in data science and convert organized weak learners into strong learners. Gradient boosting works as adaptive boosting algorithms but differ in optimize problems. It covers loss function and tries to optimize it to reduce error residuals (Table 6).

1.3.7. XG Boost Ensemble Method

Song S. et al., (2020), introduced about XG Boost sequential ensemble model in data science. It is part of Meta classifier in machine learning environment and it works with decision tree terminal nodes and calculates shrinking leaf nodes. This algorithm reduces high correlation between tree and work as gradient boosting but it has more speed and wide range compare to gradient boosting (Table 7).

Table 5. Representation of AdaBoostM1 algorithm performance on heart disease attributes

Class distributions		Class distributions	
cp <= 0.5		thalach <= 160.5	
zero	one	zero	one
0.7545271629778671	0.2454728370221328	0.5519147430473138	0.4480852569526862
cp > 0.5		thalach > 160.5	
zero	one	zero	one
0.23484848484848486	0.7651515151515151	0.3445905928029681	0.6554094071970319
cp is missing		thalach is missing	
zero	one	zero	one
0.4868292682926829	0.5131707317073171	0.4869316512173426	0.5130683487826573
Weight: 1.15		Weight: 0.34	
		Number of performed Iterations: 10	

Table 6. Representation of gradient boosting algorithm performance on heart disease attributes

```
Iteration 1
Class 1 (target=zero)
cp <= 0.5: 1.0181086519114688
cp > 0.5: -1.0606060606060606
cp is missing: -0.05268292682926829

Class 2 (target=one)
oldpeak <= 0.05: -0.5234790776842392
oldpeak > 0.05: 0.23831414928504496
oldpeak is missing: 0.005200812248078305
Number of performed iterations: 10
```

2. RELATED WORK

Cai et al., (2020), analyzed ECG recordings of 16,557 patients for atrial fibrillation detection. They used deep learning and densely connected neural network for better prediction. Authors calculated accuracy (99.35%), sensitivity of (99.19%) and specificity (99.44%).

Table 7. Representation of XG Boost algorithm performance on heart disease attributes

```
Weight: 1.01
thal <= 2.5: one
thal > 2.5: zero
thal is missing: zero

Weight: 0.48
thalach <= 160.5: zero
thalach > 160.5: one
thalach is missing: one
Number of performed Iterations: 10
```

Buettner et al., (2020), discussed 499 patients of ECG and their recording. They used random forest, spectral analysis in machine Learning. Authors calculated (96.77%) classification accuracy for ECG medical dataset.

Magesh and Swarnalatha (2020), predicted 303 samples of heart patients in machine learning. Authors used random forest, cluster-based decision tree learning and calculated (89.30%) accuracy.

Harimoorthy and Thangavelu (2020), considered support vector machine with RBK and analyzed support vector machine with linear, polynomial, random forest and decision tree. Authors calculated (98.3%), (98.7%) and (89.9%) accuracy for SVM with polynomial, random forest and decision tree.

Miled et al., (2020), considered cases of heart patients and control 11,558 heart cases from 15 and 25 different institutions. They have used random forest for dementia prediction in machine learning. Authors calculated (77.43%) classification accuracy for by random forest in heart disease.

Amin, M.S., et al., (2019), discussed about hybrid technique for better prediction in heart disease medical dataset. They used naïve bayes, logistic regression, k-NN, support vector machine, neural network and vote. Authors generate a hybrid technique with naïve bayes and logistic regression and find (87.4%) classification accuracy.

Verma AK et al. (2020), analyzed 366 instances and 34 attributes of heart disease in machine learning. They have used PAC, LDA, RNC, BNB, NB, and ETC, bagging, AdaBoost, and gradient boosting algorithms and calculated (99.68%) classification accuracy.

We have considered review work (2012 -2020) and find the various accuracy near about (99%). In the work, we have compared parallel and sequential ensemble methods to reduce above drawback in prediction. The parallel and serial ensemble methods organized by J48 algorithm, Reduced Error Pruning and Decision Stump Algorithm decision tree based algorithms. In this research work, we have tried to test four ensemble methods and finally find Random Forest ensemble method provide better result (100%) accuracy.

3. METHODOLOGY

In this stage, we have organized all the experimental setup by machine learning algorithms. The data description section describes different types heart related problems with related attributes and visualized their distribution by classifier algorithms using python tool.

3.1. Data Preparation

The heart disease medical dataset consists of 1025 individuals' instances with 14 attributes. The organized dataset collect from UCI Repository in binary and numeric format (Table 8). The target variable contain class level dependable variable in algebraic format. We calculated on the basis of class as:

target
0 → 499
1 → 526

3.1.1. Pairs Plot

Pérez-Enciso M et al., (2020), introduced about pair plot of attributes distribution. A pairs plot assigns the features distribution of single variables and strength between two variables. This method assigns relationship between two variables and identifies the trends for analysis (Figure 1).

Table 8. Representation of data description of heart disease attributes

Attributes	Description
Age	Age of the patients in years
Sex	(1 = Male; 0 = Female)
Cp	Categories chest pain (0: Typical Angina, 1: Atypical)
Trestbps	Resting blood pressure (mm hg)
Chol	Measure serum cholesterol in Mg/Dl
Fbs	(Measure fasting blood sugar > 120 Mg/Dl) (1 = True; 0 = False)
Restecg	Measure electrocardiographic (0: Normal; 1: Abnormality)
Thalach	Measure maximum heart rate
Exang	Calculated exercise induced Angina (1 = Yes; 0 = No)
Oldpeak	Measure depression induced by exercise relative to rest
Slope	Measure slope of peak exercise (0: Upsloping; 1: Flat; 2: Downsloping)
Ca	Identify major vessels (0-3) colored by Flourosopy
Thal	(1 = Normal; 2 = Fixed Defect; 3 = Reversable Defect)
Target	Diagnosis of heart disease (0: Disease; 1: Nondisease)

3.1.2. Statistical Matrix Evaluation

In this research, we have used some computational formula for Kappa Statistic, Classification Accuracy and Errors on heart disease. Yadav DC, Pal S., (Chaurasia & Pal, 2020a; Chaurasia & Pal, 2020b; Kumar et al., 2019; Verma et al., n.d.; Yadav & Pal, 2019a; Yadav & Pal, 2019b; Yadav & Pal, 2020b).

The kappa statistic measure of how closely the instances classified using expected and observed agreement in data science environment. It is formulated as:

$$\text{(Observed Values – Expected Values)/ (1 – Expected Values)} \qquad (1)$$

The strong and weak kappa statistic value varies between (0-1). When kappa calculated values tends to 1 then it will be agree near perfection otherwise no perfection:

$$\text{Accuracy} = (\text{correctly predicted class / total testing class}) \times 100\% \qquad (2)$$

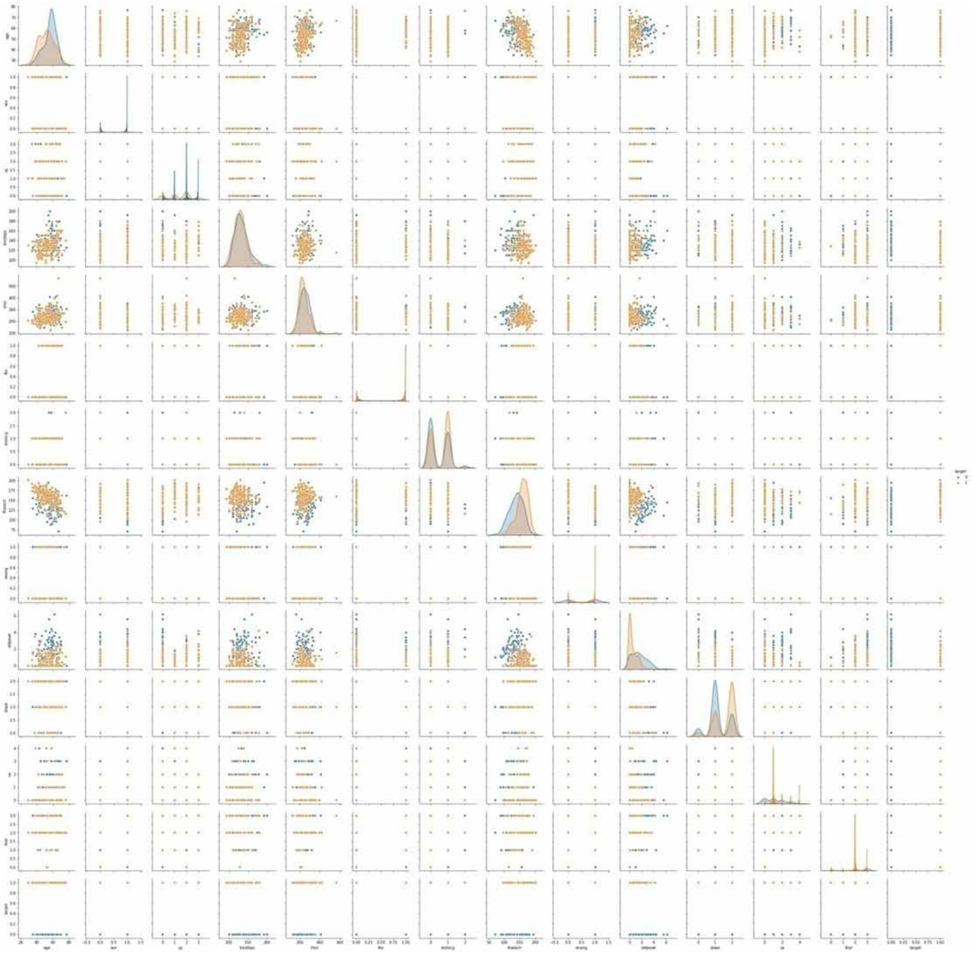
$$\text{Correctly Classified} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN}) \qquad (3)$$

where TP= true positives, FN false negatives, FP=false positives and TN=true negatives, respectively.

$$\text{Mean Absolute Error (MAE)} = 1/e \sum_{i=1}^{e} \left| b_i - b \right| \qquad (4)$$

where, e number of errors and |b_(i)-b| represents errors.

Figure 1. Representation of features distribution by pair plot of heart disease attributes



$$\text{Relative Absolute Error (RAE)} = \frac{\left[\sum_{i=1}^{e} W_{i\ -} B^{2}\right]^{1/2}}{\left[\sum_{i=1}^{e} B^{2}\right]^{1/2}} \qquad (5)$$

where W_i $\mathit{Calculate\ predicted\ value\ and}$ B calculate actual value for (i=1..e) samples:

Root Mean Square Error (RMSE) = √1/e∑_(i=1)^e 〖〖(c)〗_(i -) o_(i))〗^2 (6)

where, (e, c and o) represents variables of sample, forecasts and observed values respectively for(i=1…e).

$$\text{Root Relative Square Error (RRSE)} = \frac{\sum_{j=1}^{e} (W_{ij-} T_{ij})^2}{\sum_{j=1}^{e} (T_j - \overline{T})^2} \tag{7}$$

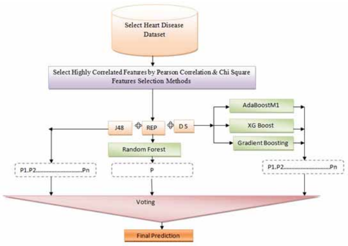
where, $W_{(ij)}$ is the value predicted by the individual program i for sample case j (out of n sample cases); T_j is the target value for sample case j; $\overline{T}$ represents perfect fit in sample cases:

$$\overline{T}\sum_{j=1}^{n}\left(T_j - T\right) \tag{8}$$

3.2. Proposed Model

In this paper, we have used Pearson Correlation and Chi-Square features selection based algorithms for heart disease attributes correlation strength. The main objective of this research to achieved highest classification accuracy with very less error. So, we have used parallel and sequential ensemble methods to reduce above drawback in prediction. The parallel and serial ensemble methods organized by J48 algorithm, Reduced Error Pruning and Decision Stump Algorithm decision tree based algorithms. We have used Random Forest ensemble method for parallel randomly selection in prediction and various sequential ensemble methods as AdaBoost, Gradient Boosting and XGBoost Meta classifiers. After the features selection trained on (75%) dataset and the test on (25%) with tree algorithms with ensemble method. The final prediction has measured by average voting algorithms (Figure 2).

Figure 2. Representation of proposed ensemble method for heart disease attributes prediction


Select Heart Disease Dataset
Select Highly Correlated Features by Pearson Correlation & Chi Square Features Selection Methods
AdaBoostM1
J48
REP
D S
XG Boost
Gradient Boosting
Random Forest
P1.P2............................Pn
P
P1.P2............................Pn
Voting
Final Prediction

4. RESULTS

In this research, we have used heart disease medical dataset consists of 1025 individuals' instances with 14 attributes and Pearson Correlation and Chi-Square features selection based algorithms for heart disease attributes correlation strength.

4.1. Pearson Correlation

Fu T et al., (2020), introduced about attributes correlation by Pearson's Correlation method. The Pearson's Correlation method generates relationship between the continuous features and the target variables. Pearson Correlation identifies feature relations to the response variable and measures linear correlation between two variables. This method measure the results by [-1;1], with -1 assign perfect negative correlation and +1 assign perfect positive correlation. The perfect negative correlation means one variable increases, the other decreases and perfect positive correlation means no linear correlation between the two variables. We have examined and find all the attributes in this heart disease are linearly correlated with maximum number of positive correlations (Figure 3).

4.2. Chi-Square

Bahassine S et al., (2020), introduced about Chi-Square feature selection for better score values of attributes. Chi-Square is a feature selection algorithm in machine learning and test the relationship between important features in dataset. The chi-square tests the independence of two events in statistics. This algorithm provides observed count values and expected count values and measures how expected

Figure 3. Representation of Pearson correlation method for heart disease attributes

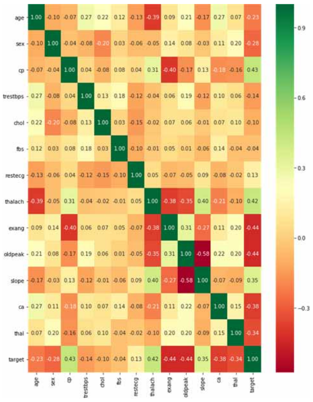


Figure 4. Representation of Chi-Square method for heart disease attributes score

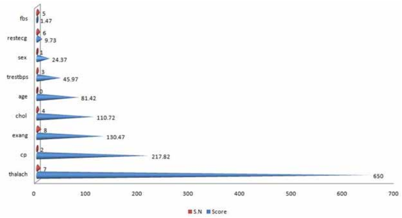


Table 9. Representation confusion matrix for heart disease attributes prediction

Confusion Matrix			
J48	**REP**	**Decision Stump**	**Random Forest**
a b <-- classified as 421 78 l a = zero 65 461 l b = one	a b <-- classified as 464 35 l a = zero 38 488 l b = one	a b <-- classified as 375 124 l a = zero 122 404 l b = one	a b <-- classified as 499 0 l a = zero 0 526 l b = one
XG Boost	**AdaboostM1**	**Gradient Boosting**	
a b <-- classified as 375 124 l a = zero 122 404 l b = one	a b <-- classified as 419 80 l a = zero 81 445 l b = one	a b <-- classified as 486 13 l a = zero 11 515 l b = one	

count with observed values. In this research, we have examined and find attribute "thalach" calculated high score and attribute "fbs" calculated low value between 9 selected positive scores (Figure 4).

In the statistical analysis used methods improved the prediction in heart disease using Pearson correlation and chi-square feature selection by eliminating the misclassified instances. We compare the performance of four ensemble methods, namely Random Forest, AdaBoostM1, Gradient Boosting and XG Boosting. The prediction model evaluated all performance by confusion matrix's true positive, false positive, true negative and false negative with class values (Table 9).

In this analysis, we examined three algorithms: J48, Reduced Error Pruning and Decision Stump and generated an ensemble method Random Forest. Random forest ensemble method calculated 100% classification accuracy, which is high compare to other three algorithms. The kappa values of random forest is high and mean absolute, root mean squared, relative absolute and root relative error values are always low of Random Forest compare to J48, Reduced Error Pruning and Decision Stump algorithms in Table10.

In the second analysis, we compared three algorithms: J48, Reduced Error Pruning and Decision Stump with three sequential ensemble methods, namely AdaBoostM1, XG Boost and Gradient Boosting. These three ensemble methods, namely AdaBoostM1, XG Boost and Gradient Boosting generated for J48, Reduced Error Pruning and Decision Stump algorithms. In this experiment, we find XG Boost calculated 98.5% classification accuracy, which is high compare to other algorithms in this table. The kappa values of XG Boost is high and mean absolute, root mean squared, relative absolute and root relative error values are always low of XG Boost compare to other algorithms in this Table 11.

Table 10. Representation of Analysis-I by J48, REP, DS and RF algorithms

Statistical Analysis	J48	Reduced Error Pruning	Decision Stump	Random Forest
Correctly Classified Instances	86.05%	92.88%	76%	100%
Incorrectly Classified Instances	13.95%	7.12%	24.00%	0.00%
Kappa statistic	0.7206	0.8575	0.5196	1
Mean absolute error	0.216	0.0951	0.365	0.0174
Root mean squared error	0.3244	0.2414	0.4275	0.0519
Relative absolute error	43.24%	19.02%	73.05%	3.48%
Root relative squared error	64.89%	48.29%	85.52%	10.38%

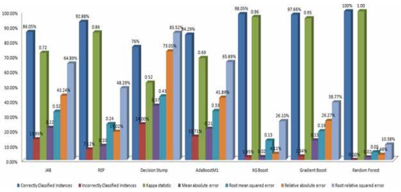
Table 11. Analysis-II by J48, REP, DS and AdaBoostM1, XG Boost and Gradient Boosting algorithms

Statistical Analysis	J48	REP	Decision Stump	AdaBoostM1	XG Boost	Gradient Boost
Correctly Classified Instances	86.05%	92.88%	76%	84.29%	98.05%	97.66%
Incorrectly Classified Instances	13.95%	7.12%	24.00%	15.71%	1.95%	2.34%
Kappa statistic	0.7206	0.8575	0.5196	0.6857	0.961	0.9531
Mean absolute error	0.216	0.0951	0.365	0.2093	0.02	0.1313
Root mean squared error	0.3244	0.2414	0.4275	0.3293	0.1304	0.1938
Relative absolute error	43.24%	19.02%	73.05%	41.89%	4.11%	26.27%
Root relative squared error	64.89%	48.29%	85.52%	65.89%	26.10%	38.77%

5. DISCUSSION

We summaries experiment I and II performances for all algorithms in Figure 5 and compare their classification accuracy and error prediction results. The figure represents four different ensemble and non-ensemble machine learning algorithms.

Figure 5. Representation of statistical analysis for accuracy and errors on heart disease attributes



During section, we compared experiment I and II and found Random forest ensemble method calculated 100% classification accuracy, which is high compare to other ensemble methods. The kappa values of random forest is high and mean absolute, root mean squared, relative absolute and root relative error values are always low of Random Forest compare to AdaBoostM1, XG Boost and Gradient ensemble methods (Table 12).

On the basis of some previous performance on different dataset authors: Yadav DC, Pal S, (Yadav & Pal, 2018; Yadav & Pal, 2019c; Yadav & Pal, 2020a), calculated high accuracy but not covered 100% accuracy on the basis of ensemble method with majority of voting. In various review paper, we have considered the high, Instances and the from 2012 to 2020 in Table 13.

All the mentioned authors used various instances and algorithms but did not covered 100% classification accuracy at very low error rate. In this experiment, we find Random Forest calculated 100% classification accuracy at low error rate.

6. CONCLUSION

In experiment, we have organized heart disease dataset from UCI repository and dataset contains total 14 attributes with 1025 instances. In whole analysis we found, Pearson Correlation generates fair matrix for heart disease features for prediction in machine learning. The feature selection technique

Table 12. Comparison of R F with AdaBoostM1, XG Boost and Gradient ensemble methods

Statistical Analysis	Random Forest	AdaBoostM1	XG Boost	Gradient Boost
Correctly Classified Instances	100%	84.29%	98.05%	97.66%
Incorrectly Classified Instances	0.00%	15.71%	1.95%	2.34%
Kappa statistic	1	0.6857	0.961	0.9531
Mean absolute error	0.0174	0.2093	0.02	0.1313
Root mean squared error	0.0519	0.3293	0.1304	0.1938
Relative absolute error	3.48%	41.89%	4.11%	26.27%
Root relative squared error	10.38%	65.89%	26.10%	38.77%

Table 13. Representation of previous year paper accuracy score

Authors	Instances	Algorithms	Accuracy	Year	Reference
Huang et al.	9800	Heartbeat classification, Independent component analysis & RR	98.35	2012	(Chen & Huang 2020)
Acharya et ai.	110,094	ECG,P CA,LDA,ICA, SVM, DWT & PNN	99.28	2013	(Martis et al., 2013)
Gabbouj et al.	100,389	ECG classification, CNN & BP	98.90	2015	(Kiranyaz et al., 2015)
Salim et al.	110,094	ECG, NN, LF, PCA, NFC & SVM	98.90	2016	(Elhaj et al., 2016)
Li et al.	90808	Heartbeat classification, Weighted RR & SVM	98.46	2017	(Chen et al., 2017)
Hagiwara et al.	109949	CNN, Deep learning & Electrocardiogram signals	94.47	2017	(Acharya et al., 2017)
Yildirim et al.	7376	LSTM, RNN, Deep learning & ECG signals	99.39	2018	(Yildirim 2018)
Baloglu et al.	100,022	ECG compression, Deep learning, Auto encoders &LSTM	99.23	2019	(Yildirim et al., 2019)
Zhao et al.	100630	CNN & electrocardiogram (ECG)	99.06	2020	(Wang et al., 2020)

namely, Chi-Square generates valuable score with corresponding features. Both features selection techniques provide help in prediction for applied machine algorithms. In this paper, all experiment divides into two parts, the first part deals with J48, Reduced Error Pruning and Decision Stump and generated a Random Forest ensemble method. This parallel ensemble method calculated high classification accuracy 100% with low error. The second part of experiment deals with J48, Reduced Error Pruning and Decision Stump with three sequential ensemble methods, namely AdaBoostM1, XG Boost and Gradient Boosting. The XG Boost ensemble method calculated better results or high classification accuracy and low error compare to AdaBoostM1 and Gradient Boosting ensemble methods. The XG Boost ensemble method calculated 98.05% classification accuracy, but Random Forest ensemble method calculated high classification accuracy 100% with low error. With the results, finally we found, Random Forest performed better results compare to all other applied machine learning algorithms in this paper.

For the future work planning, we will observe other features selection method with fuzzy, artificial and Neural Network Hybrid Ensemble Method.

ACKNOWLEDGMENT

The author is grateful to Veer Bahadur Singh Purvanchal University Jaunpur, Uttar Pradesh, for providing financial support to work as Post Doctoral Research Fellowship.

REFERENCES

Acharya, U. R., Oh, S. L., Hagiwara, Y., Tan, J. H., Adam, M., Gertych, A., & San Tan, R. (2017). A deep convolutional neural network model to classify heartbeats. *Computers in Biology and Medicine*, *89*, 389–396. doi:10.1016/j.compbiomed.2017.08.022 PMID:28869899

Amin, M. S., Chiam, Y. K., & Varathan, K. D. (2019). Identification of significant features and data mining techniques in predicting heart disease. *Telematics and Informatics*, *36*, 82–93. doi:10.1016/j.tele.2018.11.007

Bahassine, S., Madani, A., Al-Sarem, M., & Kissi, M. (2020). Feature selection using an improved Chi-square for Arabic text classification. *Journal of King Saud University-Computer and Information Sciences*, *32*(2), 225–231. doi:10.1016/j.jksuci.2018.05.010

Buettner, R., Beil, D., Scholtz, S., & Djemai, A. (2020, January). Development of a machine learning based algorithm to accurately detect schizophrenia based on one-minute EEG recordings. *Proceedings of the 53rd Hawaii International Conference on System Sciences*. doi:10.24251/HICSS.2020.393

Cai, W., Chen, Y., Guo, J., Han, B., Shi, Y., Ji, L., Wang, J., Zhang, G., & Luo, J. (2020). Accurate detection of atrial fibrillation from 12-lead ECG using deep neural network. *Computers in Biology and Medicine*, *116*, 103378. doi:10.1016/j.compbiomed.2019.103378 PMID:31778896

Catherine, O. O. (2020). Lower Respiratory Tract Infection Clinical Diagnostic System Driven by Reduced Error Pruning Tree (REP Tree). *Am J Compt Sci Inform Technol*, *8*(2), 53.

Chaurasia, V., & Pal, S. (2020a). Applications of Machine Learning Techniques to Predict Diagnostic Breast Cancer. *SN Computer Science*, *1*(5), 1–11. doi:10.1007/s42979-020-00296-8

Chaurasia, V., & Pal, S. (2020b). Skin Diseases Prediction: Binary Classification Machine Learning & Multi Model Ensemble Techniques. *Indian Journal of Public Health Research & Development*, *11*(1), 737–742. doi:10.37506/v11/i1/2020/ijphrd/193913

Chen, Q., & Huang, L. (2020). Research on Prediction Model of Gas Emission Based on Lasso Penalty Regression Algorithm. In *Artificial Intelligence in China* (pp. 165–172). Springer. doi:10.1007/978-981-15-0187-6_19

Chen, S., Hua, W., Li, Z., Li, J., & Gao, X. (2017). Heartbeat classification using projected and dynamic features of ECG signal. *Biomedical Signal Processing and Control*, *31*, 165–173. doi:10.1016/j.bspc.2016.07.010

Çınar, A., Ince, E., Gezer, M., & Yılmaz, Ö. (2020). Machine learning algorithm for grading open-ended physics questions in Turkish. *Education and Information Technologies*, 1–24.

Elhaj, F. A., Salim, N., Harris, A. R., Swee, T. T., & Ahmed, T. (2016). Arrhythmia recognition and classification using combined linear and nonlinear features of ECG signals. *Computer Methods and Programs in Biomedicine*, *127*, 52–63. doi:10.1016/j.cmpb.2015.12.024 PMID:27000289

Fu, T., Tang, X., Cai, Z., Zuo, Y., Tang, Y., & Zhao, X. (2020). Correlation research of phase angle variation and coating performance by means of Pearson's correlation coefficient. *Progress in Organic Coatings*, *139*, 105459. doi:10.1016/j.porgcoat.2019.105459

Gupta, A., Suri, B., Kumar, V., & Jain, P. (2020). Extracting rules for vulnerabilities detection with static metrics using machine learning. *International Journal of System Assurance Engineering and Management*, 1-12.

Harimoorthy, K., & Thangavelu, M. (2020). Multi-disease prediction model using improved SVM-radial bias technique in healthcare monitoring system. *Journal of Ambient Intelligence and Humanized Computing*, 1–9. doi:10.1007/s12652-019-01652-0

Kaandorp, M. L., & Dwight, R. P. (2020). Data-driven modelling of the Reynolds stress tensor using random forests with invariance. *Computers & Fluids*, *202*, 104497. doi:10.1016/j.compfluid.2020.104497

Kiranyaz, S., Ince, T., & Gabbouj, M. (2015). Real-time patient-specific ECG classification by 1-D convolutional neural networks. *IEEE Transactions on Biomedical Engineering*, *63*(3), 664–675. doi:10.1109/TBME.2015.2468589 PMID:26285054

Kumar, V., Verma, A. K., & Yadav, D. C. (2019). A Review on Solid Lipid Nano Particles (SLNS). *Research Journal of Pharmacy and Technology*, *12*(11), 5605–5613. doi:10.5958/0974-360X.2019.00971.5

Ma, B., Meng, F., Yan, G., Yan, H., Chai, B., & Song, F. (2020). Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data. *Computers in Biology and Medicine*, *121*, 103761. doi:10.1016/j.compbiomed.2020.103761 PMID:32339094

Magesh, G., & Swarnalatha, P. (2020). Optimal feature selection through a cluster-based DT learning (CDTL) in heart disease prediction. *Evolutionary Intelligence*, 1–11. doi:10.1007/s12065-019-00336-0

Martis, R. J., Acharya, U. R., & Min, L. C. (2013). ECG beat classification using PCA, LDA, ICA and discrete wavelet transform. *Biomedical Signal Processing and Control*, *8*(5), 437–448. doi:10.1016/j.bspc.2013.01.005

Miled, Z. B., Haas, K., Black, C. M., Khandker, R. K., Chandrasekaran, V., Lipton, R., & Boustani, M. A. (2020). Predicting dementia with routine care EMR data. *Artificial Intelligence in Medicine*, *102*, 101771. doi:10.1016/j.artmed.2019.101771 PMID:31980108

Palad, E. B. B., Burden, M. J. F., Torre, C. R. D., & Uy, R. B. C. (2020). Performance evaluation of decision tree classification algorithms using fraud datasets. *Bulletin of Electrical Engineering and Informatics*, *9*(6), 2518–2525. doi:10.11591/eei.v9i6.2630

Pérez-Enciso, M., Ramírez-Ayala, L. C., & Zingaretti, L. M. (2020). SeqBreed: A python tool to evaluate genomic prediction in complex scenarios. *Genetics, Selection, Evolution.*, *52*(1), 1–9. doi:10.1186/s12711-020-0530-2 PMID:32039696

Song, S., Fei, C., & Xia, H. (2020). Lithium-Ion Battery SOH estimation based on XGBoost algorithm with accuracy correction. *Energies*, *13*(4), 812. doi:10.3390/en13040812

Verma, A. K., Pal, S., & Kumar, S. (n.d.). Prediction of Different Classes of Skin Disease Using Machine Learning Techniques. In *Smart Innovations in Communication and Computational Sciences* (pp. 91–100). Springer. doi:10.1007/978-981-15-5345-5_8

Verma, A. K., Pal, S., & Kumar, S. (2020). Prediction of skin disease using ensemble data mining techniques and feature selection method—A comparative study. *Applied Biochemistry and Biotechnology*, *190*(2), 341–359. doi:10.1007/s12010-019-03093-z PMID:31350666

Virani, S. S., Alonso, A., Benjamin, E. J., Bittencourt, M. S., Callaway, C. W., Carson, A. P., & Djousse, L. et al. (2020). Heart disease and stroke statistics—2020 update: A report from the American Heart Association. *Circulation*, *141*(9), E139–E596. doi:10.1161/CIR.0000000000000757 PMID:31992061

Wang, H., Shi, H., Chen, X., Zhao, L., Huang, Y., & Liu, C. (2020). An Improved Convolutional Neural Network Based Approach for Automated Heartbeat Classification. *Journal of Medical Systems*, *44*(2), 35. doi:10.1007/s10916-019-1511-2 PMID:31853698

Yadav, D. C., & Pal, S. (2018, October). A Fair Knowledge of Bureau Report by Data Mining Algorithms. In *2018 3rd International Conference on Communication and Electronics Systems (ICCES)* (pp. 1024-1028). IEEE.

Yadav, D. C., & Pal, S. (2019a). Thyroid prediction using ensemble data mining techniques. *International Journal of Information Technology*, 1-11.

Yadav, D. C., & Pal, S. (2019b). Calculating diagnose odd ratio for thyroid patients using different data mining classifiers and ensemble techniques. *International Journal of Advanced Trends in Computer Science and Engineering*, *9*(4), 5463–5470. doi:10.30534/ijatcse/2020/186942020

Yadav, D. C., & Pal, S. (2019c). To generate an ensemble model for women thyroid prediction using data mining techniques. Asian Pacific journal of cancer prevention. *APJCP*, *20*(4), 1275. PMID:31031212

Yadav, D. C., & Pal, S. (2020b). Prediction of thyroid disease using decision tree ensemble method. *Human-Intelligent Systems Integration*, 1-7.

Yadav, D. C., & Pal, S. (2020a). Prediction of Heart Disease Using Feature Selection and Random Forest Ensemble Method. *International Journal of Pharmaceutical Research*, *12*(4).

Yildirim, Ö. (2018). A novel wavelet sequence based on deep bidirectional LSTM network model for ECG signal classification. *Computers in Biology and Medicine*, *96*, 189–202. doi:10.1016/j.compbiomed.2018.03.016 PMID:29614430

Yildirim, O., Baloglu, U. B., Tan, R. S., Ciaccio, E. J., & Acharya, U. R. (2019). A new approach for arrhythmia classification using deep coded features and LSTM networks. *Computer Methods and Programs in Biomedicine*, *176*, 121–133. doi:10.1016/j.cmpb.2019.05.004 PMID:31200900

Dhyan Chandra Yadav, (PhD) Post Doc., received his MCA from Veer bahadur Singh Purvanchal University, jaunpur in 2008, Ph.D. degrees in Computer Application from SV University Gajraulla, Amroha JapiNagar, U.P. in 2016, and working in Veer bahadur Singh Purvanchal University, Jaunpur as Post Doctoral Fellow. His research interests include pattern recognition, statistical image processing and in knowledge discovery. Dr. Yadav has authored/co-authored 25 publications in various high impact factor, peer-reviewed, journals. Dr. Yadav has written 02 book chapters and he presented about 04 papers in international conferences. Dr. Dhyan Chandra Yadav served as Conference reviewer in various conferences and in Journals.

Saurabh Pal (PhD), Head Dept. of Computer Applications., in Veer bahadur Singh Purvanchal University, Jaunpur. His research interests include machine learning, pattern recognition, software engineering, cloud computing etc. Dr. Pal has authored/co-authored 80 publications in various high impact factor, peer-reviewed, journals. Dr. Pal has written 25 book chapters and he presented about 30 papers in international conferences. Dr. Saurabh Pal served as reviewer in various conferences and in Journals.