COMPARISON OF ADABOOST AND BAGGING ENSEMBLE METHOD FOR PREDICTION OF HEART DISEASE

Yusuf Olatunde [1], Lawrence Omotosho [2], Caleb Akanbi [2]

[1] Summit University, Offa, Kwara State, - Nigeria, Department of Computer Science
[2] Osun State University Osogbo Campus, Osun State – Nigeria, Department of Computer Science

Corresponding Author: Caleb Akanbi, akanbico@uniosun.edu.ng

ABSTRACT: The medical industry arguably generates the largest amount of data on a daily basis. Extraction of new and useful information from the bulk of data generated is very tedious. Although, it contributes to the quality of service rendered in the health sector. Data mining techniques are among the major approach that shows promising result when applied in diagnosing patient and prediction of diseases. In this study, AdaBoost and Bagging are used to support classifiers such as Naïve Bayes, Neural Network in prediction of heart disease while Random Forest was applied separately. Comparison of the experiment results focus majorly on the ensemble method used (AdaBoost and Bagging). With respect to this study, Bagging outperforms AdaBoost in term of Accuracy and other parameters such as Kappa Statistics, weighted average of ROC, Precision and MCC. It is therefore recommended as a good supportive technique for weak classifiers. Although, both Bagging and AdaBoost decline in performance when applied on rigorous dataset.
KEYWORDS: Heart Disease, Adaptive Boosting, Bootstrap Aggregation, Neural Network, Naïve Bayes, Random Forest

1. INTRODUCTION

Medical practitioners have labelled diseases such skin disease, breast cancer, lung cancer, leukaemia (cancer of the blood), heart disease, ebola virus, monkey pox and many more as a deadly disease that requires urgent attention ([CS12], [JTS16]). Experts encounter a lot of difficulties in reaching conclusion on several deadly diseases due to symptoms similarities ([VB17]). Technologically, Data mining techniques are employed to support the effort of medical personnel in reaching early conclusion and making accurate decision in order to save the lives of the patients ([VB17]). These techniques require improvement because the rate of the death caused by previously known and newly identified disease is currently high, most especially in the developing countries. Therefore, this study concentrates on prediction of heart disease. The content of this paper is divided into 5 sections with subsections. The subsections under section 1 (that is, 1.1 to 1.6) describes the human heart, it functions and related diseases, follow by the data mining technique, ensemble methods and other classifiers used for experiment in this study. Section 2 presents review of the related work. Section 3 discuss the research methodology, source and description of data. In section 4, result of the study is presented follow by the summary and conclusion in section 5 which also captures the recommendation of this study.

1.1 HUMAN HEART AND ITS ASSOCIATED DISEASES

The heart is an essential organ of our body. The improper functioning of the heart affects other body parts such as the kidney and brain. Inefficient circulation of blood in the body makes organs like brain suffers. Life is extremely dependent on the efficient working of the heart because once the heart stops works the end (death) occurs. Heart disease is sequence of contrary conditions that affect the heart. Heart disease is one of the deadly diseases in human being and can lead to sudden death ([CS12], [CB15], [L+17]).
The heart shown in Figure 1 is responsible for the production of oxygenated blood and distribute it to all part of the body through the arteries. All deoxygenated blood in the body are transported to the heart by the vein for purification before it can be useful to the body. A normal and healthy heart should beat 72 times per minute and more than hundred thousand times pumping approximately two thousand gallons of blood throughout the body every day. Automatically, malfunctioning of the heart affects other parts of the body such as kidney and brain. Some of the diseases associated to the heart are cardiovascular disease ([L+17]), cardiomyopathies ([K+14]), valvular heart disease ([V+12]), pericardial disease, congestive heart failure ([P+16]) and coronary artery disease.

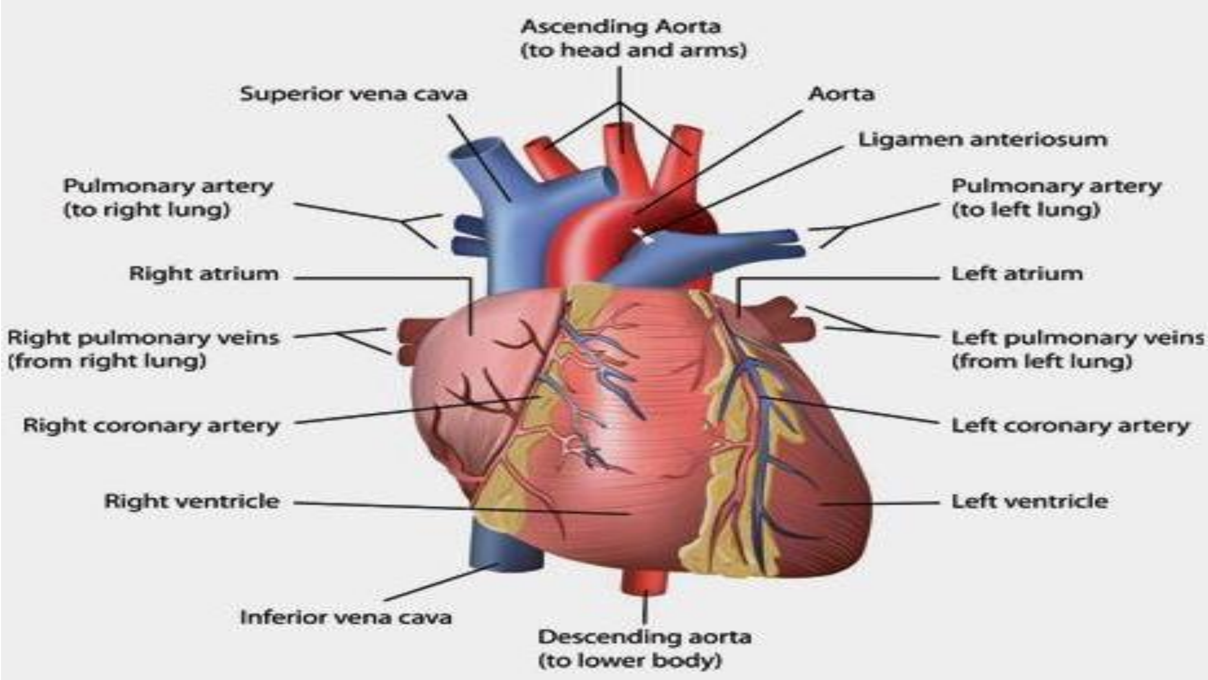

Fig. 1: The Human Heart

1.2 DATA MINING TECHNIQUES

Data mining is becoming one of the most indispensable and inspiring area of research because of its ability to discover significant information from a pool of data ([SS16]). The extracted information can be usefully applied in health data analysis and science exploration, production control, market analysis, and business management to emerging designs, ([RP16]). Different techniques are used in data mining to extract needful information such as Classification, Clustering, Association and so on. In this paper, meta-algorithm such as Bagging and AdaBoost are applied on weak classifiers that is, Naïve Bayes (NB), Artificial Neural network (ANN) to improve their performance on prediction of heart disease. RF is considered in this study to be an application of Bagging ensemble method on Decision Tree (DT). The output is then compared based on Accuracy, Kappa Statistic (KS), Mathews Correlation Coefficient (MCC), Receiver Operating-characteristic Curve (ROC) Area and Precision.

1.3 ENSEMBLE METHODS

In machine learning, ensemble methods entail the use of different learning algorithms to acquire better predictive performance compare to what is obtainable form a single algorithm. Therefore, it is a set of classifiers that learn a target function and their respective predictions are joined to classify new examples. It generally increases the performance of a set of classifiers on an area. It is widely adopted due to the fact that it can incorporate large volumes of information, diminishes the danger of choosing an ineffectively performing classifier and requires little information. Some of the common ensemble methods are Bayes optimal classifier, Bootstrap Aggregating (Bagging), Adaptive Boosting (AdaBoost), Random forest, Bayesian parameter averaging, Bayesian model combination, Bucket of models, Stacking and voting ([N+16]).

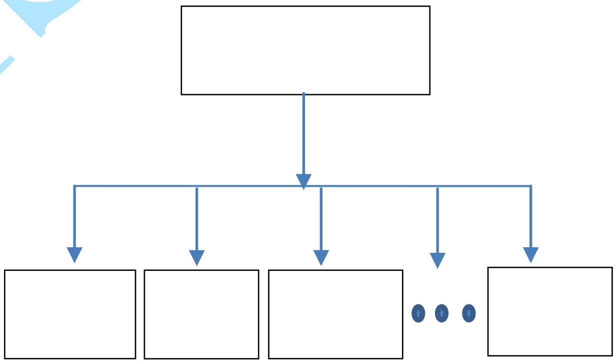

Fig. 2: Common examples of ensemble methods

1.4 ADAPTIVE BOOSTING

Adaptive Boosting is also referred to as "AdaBoost". It was first described by Yoav Freund and Robert Schapire in 1995 as a meta-algorithm in machine learning ([Sch13]). Its primary objective is to support another machine learning algorithm to become more effective. The result of other algorithms that is, weak classifier or weak learners (WL) are collated as a weighted sum which denotes the concluding result of the classifier that is boosted ([Jaa16]). It is adaptive because successive weak learners are tweaked to favour the misclassified instances by former classifiers. The general concept of AdaBoost is shown in Figure 3 ([Sin10]).

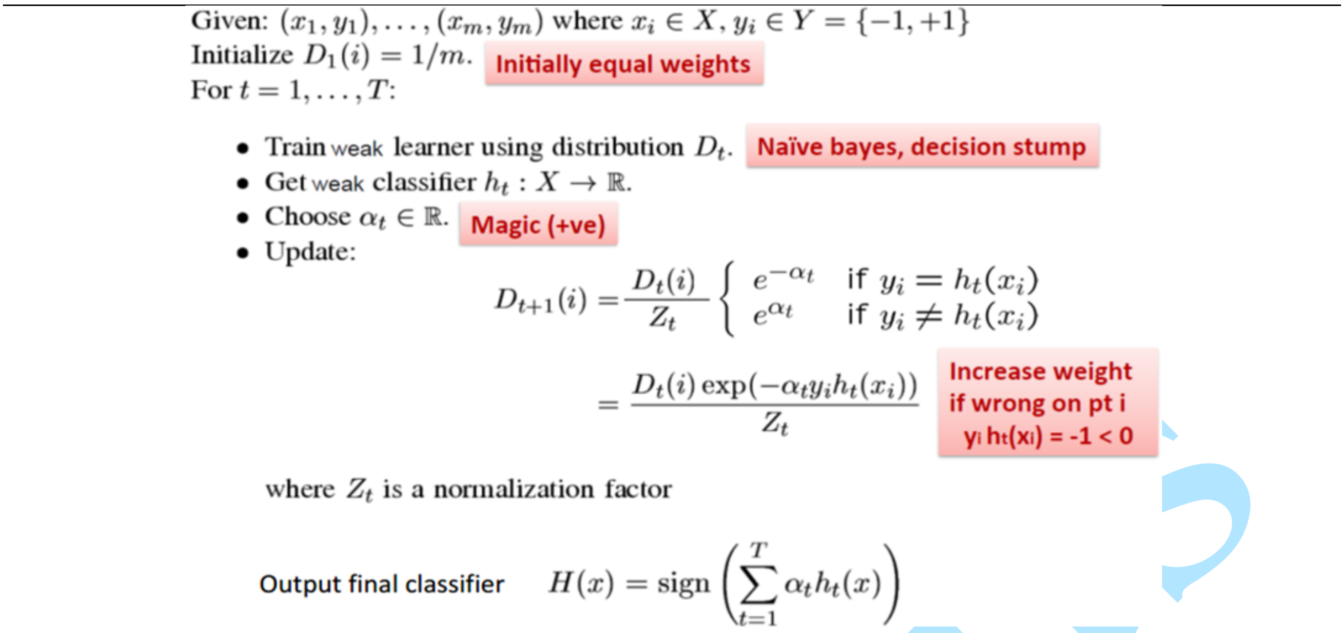

Fig. 3: AdaBoost Ensemble Algorithm

It can be understood form Figure 3 that: given m numbers of training examples that is (x_1, y_1) to (x_m, y_m) where the x_i's exist in X, and the labels $y_i = \{-1, +1\}$. where t = 1,2,3,4 to T, for each value of t, a distribution D_t is computed as shown in Figure 2 over the numbers of examples, and a particular WL algorithm is used to find a weak hypothesis h_t: X = {−1, +1}, where the WL aim is to find a weak hypothesis with low weighted error e^t that is relative to D_t. The collated hypothesis H computes the sign of a weighted combination of weak hypotheses using equation 1.

$$H(x) = \sum_{t=1}^{T} \alpha_t h_t(x) \qquad (1)$$

Therefore, H computes the weighted majority vote of the weak hypotheses h_t that is, WL where each is assigned a particular weight.

1.5 BOOTSTRAP AGGREGATING

Bootstrap Aggregating termed as "Bagging" is also an ensemble method presented in 1994 by Breiman to enhance classification by joining the classifications of arbitrarily generated training sets. It reduces variance, prevent overfitting, enhance steadiness and accuracy of machine learning algorithm. It can be applied on any classification or regression method but intentionally developed for any classification and often applied on decision tree. Bagging utilize more than one variant of a training set by using bootstrap, that is, sample with replacement. Each of this dataset is use to train another model. The result of each model is joined by averaging (for regression) or voting (for classification) to generate only one output. Bagging is only effective when using unstable nonlinear models ([Bre96]). The operation involve in Bagging is shown in Figure 4.

1.6 RANDOM FOREST

Random forest is an ensemble method that focus on improving DT. It collates all the results of several DT. Each of the tree is created based on a bootstrap sample (which is entail in Figure 5) and the values of a particular set of random vectors. A fixed probability distribution is use to generate random vectors ([Ens16]). Figure 5 diagrammatically illustrate the concept of random forest. T* in Figure 5 is the combine output of several decision trees and it serves as the final step of random forest to generate it output.

Bagging Ensemble Algorithm
Parameter
A → prediction algorithm
Y → Original dataset
K → number of training instances in Y
Y' → generated dataset from Y
T → iterations

Bootstrap sampling (Y)
initialize Y' as an empty set
while K > 0:
randomly select an instance j from Y //sampling with replacement
add j to Y'
K--
return Y'

Bagging (A, Y, T)

generate model
for i = 1, 2, 3, 4, … to….. T
compute a bootstrap sample Y(i) from Y
let R(i) be the outcome of training A on Y(i)
prediction of a particular test instance such as j
for i = 1, 2, 3, 4, … to….. T
let Bg(i) = outcome of R(i) on j

return the most appeared class among Bg(1), Bg(2) , Bg(3), …. to … C(T)

Fig. 4: Bagging Ensemble Algorithm

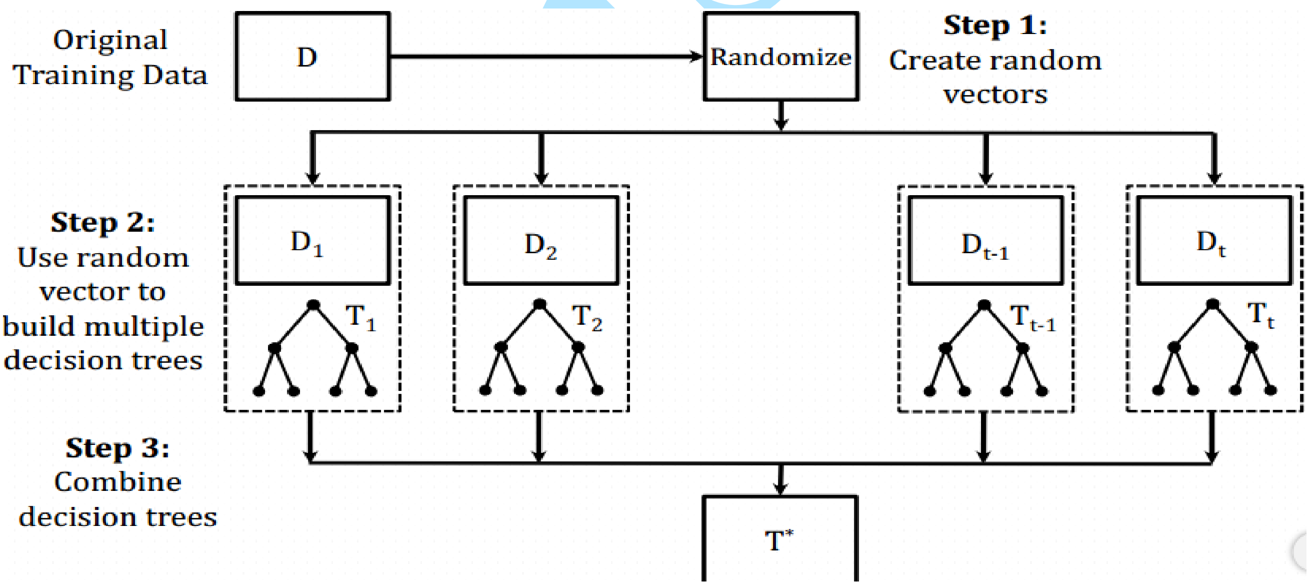

Fig. 5: Concept of Random Forest

1.7 NAÏVE BAYES

Naive Bayes sometimes call Bayes Rule is based on Bayes' Theorem and recognised as the origin of several data mining and machine-learning approaches. The technique is use in creating models with predictive competences. NB is a classifier method mostly used when the dimensionality of the input is high ([RP16]). Although, it belongs to the family of probabilistic classifiers, it is simple and easy to understand algorithm, nevertheless it can often outperform more sophisticated classification algorithm ([VB17]). It generally assumes that features in a class are present independently. The probability concept used in NB is stated in the equation below.

$$\underbrace{P(c|x)}_{A} = \frac{\overbrace{P(x|c)}^{B}\,\overbrace{P(c)}^{C}}{\underbrace{P(x)}_{D}} \qquad (2)$$

$$P(c|X) = P(x_1|c) * P(x_2|c) * P(x_3|c) * \ldots\ldots * P(x_n|c) * P(c) \qquad (3)$$

In equation 2:

A denotes posterior probability of *class* (*target*, c) in the presence of *predictor* (*attributes*, x).

C denotes *class* prior probability.

B denotes possibility that is, the probability of *predictor* given *class*.

D denotes prior probability of *predictor*.

1.8 ARTIFICIAL NEURAL NETWORK

The objective of the ANN is to tackle all difficulties in a similarly way that the human brain would but neural systems are more abstract. Recently, projects with ANN usually compose of neural units that are within thousands and millions in number with millions of connections. Recent studies on heart generally employ unique patters within the neural systems ([SS14]). Example is using deeper or complex connections and link the processing layers instead of localizing the gathering of neurons ([RPJ17], [SS16]). Figure 6 shows the diagrammatic representation of ANN concept.

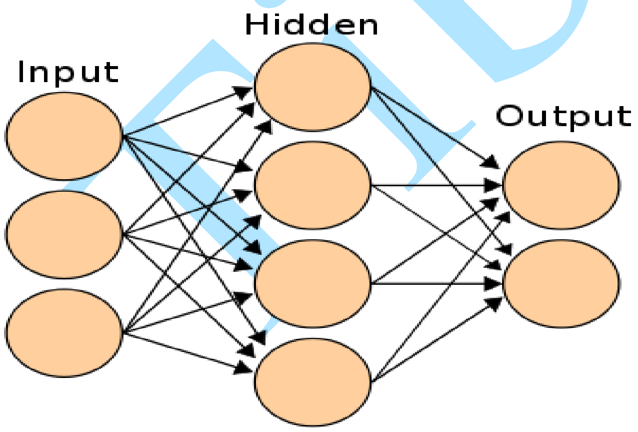

Fig. 6: Concept of Artificial Neural Network

2. RELATED WORKS

([VS13]) finds the best classifier for predicting and diagnosing heart diseases. Three different data mining algorithms was used namely Iterative Dichotomized 3 (ID3), Classification and Regression Tree (CART), and Decision Table (DTa) extracted from a DT or rule-based classifier to develop the prediction model using 10-fold cross validation was used to estimate the unbiased nature of the data. The experiment revealed that CART performs better in terms of accuracy when compared to ID3 and DTa.

([LJ16]), employs ANN to predict heart diseases and its risk at early stage. The study considered factors such as Personal information, Medical history, Diet and Lifestyle for prediction. The study shows that the system is 70% accurate.

([Kee17]) compares the performance of Decision tree (J48), Random forest and Naïve Bayes classification on prediction of heart disease. The algorithms were implemented using WEKA API. The prediction was done based on the patient's medical records such as age, blood pressure, sex, cholestesrol and blood sugar. The study revealed that NB classifier gives low error rate and high precision while RF performs better than J48 DT.

([VB17]), worked on predicting heart disease using Naïve Bayes as classification algorithm and Laplace Smoothing as Smoothing technique to develop a Decision Support System. The Decision Support System was tested using 6 attributes and 13 attributes. The study shows to be more accurate when 13 attributes was used.

([CS12]) also worked on predicting heart disease with data mining techniques. The study adds obesity and smoking to 13 attributes used in previous research. Naïve Bayes, Decision Trees, and Neural Network are the techniques applied on heart disease database. The study reveals based on accuracy that Neural Networks, Decision Trees, and Naive Bayes gives 100%, 99.62%, and 90.74% respectively.

([JTS16]) explain that death is one of the risks associated to heart disease. The study compares some techniques that utilize the general principles of Decision Tree. The algorithms are J48, Logistic Model Tree (LMT) and Random Forest. The dataset used consists of 303 instances and 76 attributes. At the end of the study's experiment, J48 tree gives the highest accuracy with 56.76% compared to LMT and Random Forest. Therefore, better algorithms or hybrid techniques are need to be develop.

([SR17]) used the heart disease collated by Cleveland, Switzerland, Hungarian and Long Beach for experiment in a study that focus on coronary heart disease. The data mining tools used are WEKA tool and Rapid Miner tool. While considering C4.5, Random Forest (RF), Random Tree (RT), REP Tree, Naïve Bayes, MLP, Support Vector Machine (SVM) algorithm, WEKA tool gives the highest accuracy with SVM achieving 66.67% with Hungarian data set while Decision Stump gives accuracy of 63.94% with same Hungarian data and Rapid miner tool.

3. METHODOLOGY

This study compares the performance of ensemble methods (Bagging and AdaBoost) by applying them on weak classifier Naïve Bayes and ANN in prediction of heart disease. That is, this study entails comparison of:

i. Bagging technique applied on Naïve Bayes
ii. Bagging technique applied on ANN
iii. Random Forest (considered as, Bagging apply on Decision Tree)
iv. AdaBoost technique applied on Naïve Bayes
v. AdaBoost technique applied on ANN

Five open source datasets of heart disease are used and 10-fold cross validation was used to estimate the unbiased nature of the data. The description of dataset used is presented in section 3.1. Table 1 shows each of the dataset are collated from different source and contains different number of instances. WEKA tool is use to perform the experiment. The algorithms (that is, weak classifiers) used in this study are selected based on the report of studies that focus on prediction of heart disease. The essence of using more than one dataset in this study is to investigate the reliability of the technique when there is large volume of data. Section 3.2 explain the metrics used in comparing the performance of the algorithms.

3.1 DESCRIPTION AND SOURCE OF DATASET

In this study, all the heart disease dataset used for experiment are described in Table 1. Out of 79 attributes present in each dataset, only 16 shown in Table 1 are considered during the experiment. That is, two additional attributes are added to the wildly used 14 attributes. That is, "Painloc" (chest pain location) and "Smoke" (if the patient is a smoker or not). This is to ensure significant observation of the ensemble techniques used.
Example of an instance in the dataset corresponding to the attributes shown in Table 1 is given below.
"63, male, 1, typ_angina, 145, 233, no, t, left_vent_hyper, 150, no, 2.3, down, 0, fixed_defect, '<50'**"**.

Table 1: Source of heart disease dataset

S/N	DATASET NAME	COLLATED FROM	COLLATED BY	ATTRIBUTE USED	INSTANCE	SOURCE (OPEN)
1.	Switzerland	University Hospital, Zurich, Switzerland	William Steinbrunn, M.D.	i. Age ii. Sex iii. Painloc iv. Cp v. Trestbps vi. Chol vii. Smoke viii. Fbs ix. Restecg x. Thalach xi. Exang xii. Oldpeak xiii. Slope xiv. Ca xv. Thal xvi. Num	123	http://archive.ics.uci.edu/ml/datasets/heart+Disease or http://tunedit.org/repo/UCI/heart-statlog.arff
2.	Long Beach VA	V.A. Medical Center, Long Beach	Robert Detrano, M.D., Ph.D.		200	
3.	Statlog	University Hospital, Basel, Switzerland:	Matthias Pfisterer, M.D.		270	
4.	Hungary	Hungarian Institute of Cardiology. Budapest	Andras Janosi, M.D.		294	
5.	Cleveland	Cleveland Clinic Foundation	Robert Detrano, M.D., Ph.D.		303	

3.2 PERFORMANCE EVALUATION METRIC

The confusion matrix shown in Figure 7 contains True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). They are used to illustrate the derivation of considered metrics used in this study. The metrics considered for performance metrics are Accuracy (Acc), Kappa Statistic (KS), Mathews Correlation Coefficient (MCC), Receiver Operating-characteristic Curve (ROC) Area and Precision (Pr). Table 3 show how each of the metrics are obtained.

Table 2: Description of Performance Metrics

S/N	METRIC	DESCRIPTION
1.	Accuracy (%)	$\frac{Number\ of\ Correct\ Prediction}{Total\ Number\ of\ Prediction} * 100$ That is, $\frac{TP+TN}{TP+TN+FP+FN} * 100$
2.	KS	$W = \frac{TP + FN}{TP + TN + FP + FN}$ $X = \frac{TP + FP}{TP + TN + FP + FN}$ $Y = \frac{FP + TN}{TP + TN + FP + FN}$ $Z = \frac{FN + TN}{TP + TN + FP + FN}$ $KS = \frac{Accuracy - ((W * X) + (Y * Z))}{1 - ((W * X) + (Y * Z))}$
3.	MCC	$\frac{(\mathrm{TP} * \mathrm{TN}) - (\mathrm{FP} * \mathrm{FN})}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}}$
4.	ROC	$\frac{TP}{2(TP + FP)} + \frac{TN}{2(TN + FN)}$
5.	Precision	$\frac{TP}{TP + FP}$

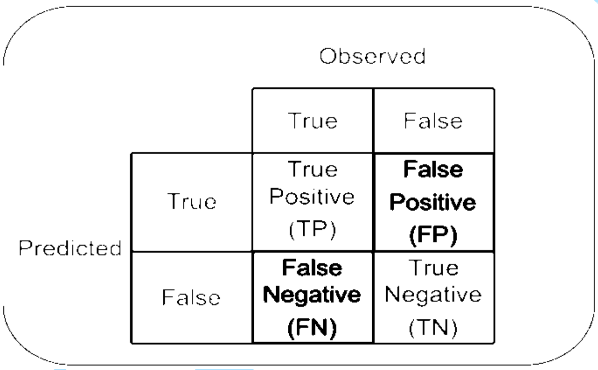

Fig. 7: Confusion Matrix

4. RESULT AND DISCUSSION

The accuracy of the techniques considered in this study fluctuate as the size of the database is changed. Table 3a and Table 3b show the performance of the considered techniques in terms of accuracy based on different dataset used. In Table 3a and Table 3b, CC denote Correctly Classified and IC denote Incorrectly Classified. It also captures the output of the experiment when RF is used and when NB, ANN are both supported by Bagging and Adaboost.

For the dataset with 123 instances shown in Table 3a, Bagging outperform both Adaboost and RF with 42.3% accuracy. Adaboost gives 36.5% when used to support NB and RF gives 34.9% accuracy. It should be noted that the result obtained at the end of the experiment with Switzerland dataset is relatively poor and thus, cannot be use to recommend a technique for the health sector.

Considering the Long Beach dataset with 200 instances shown in Table 3a, Bagging also takes the lead with 37.5% accuracy compared to Adaboost and RF which gives 33.5% and 33% accuracy respectively. For the Long Beach dataset, the accuracy is obtained for all the techniques is not convincing enough to improve medical treatment.

The third dataset labelled Statlog contains 270 instances. Experiment with this dataset produce an impressive result. Table 3a shows that Bagging maintain it led with 83.3% accuracy compared to 82.6% and 81.0% produced by Adaboost and RF respectively. There is some level of reliability in the techniques when used on the Statlog dataset considering the level of accuracy obtained. Although there is room for improvement.

Table 3a: Comparison in terms of Accuracy

Techniques		123 Instances			200 Instances			270 Instances		
		CC	IC	Acc (%)	CC	IC	Acc (%)	CC	IC	Acc (%)
Random Forest		43	80	34.9	66	134	33	220	50	81.0
Bagging	Naïve Bayes	48	75	39.0	59	141	29.5	225	45	83.3
	ANN	52	71	42.3	75	125	37.5	217	43	80.4
AdaBoost	Naïve Bayes	45	78	36.5	67	133	33.5	223	47	82.6
	ANN	38	85	30.8	32	138	31	215	55	79.6

Table 3b: Comparison in terms of Accuracy

Techniques		294 Instances			303 Instances		
		CC	IC	Acc (%)	CC	IC	Acc (%)
Random Forest		239	55	81.3	170	133	56.1
Bagging	Naïve Bayes	237	57	80.6	176	127	58.1
	ANN	233	61	79.3	169	134	55.8
AdaBoost	Naïve Bayes	236	58	80.2	169	134	55.8
	ANN	227	67	77.2	169	134	55.8

Table 3b show the tabular comparison of the considered techniques in terms of accuracy based on Hungary and Cleveland dataset which have 294 and 303 instances respectively. For Hungary dataset, the output of the experiment shows that RF slightly surpasses Bagging and Adaboost with 81.3% accuracy while Bagging and Adaboost almost perform at the same level with 80.6% and 80.2% accuracy respectively. At this point, this study observes a slight drop in the performance Bagging and Adaboost compared to the result obtained with the Statlog dataset.

Experiment with Cleveland dataset gives an accuracy of 58.1%, 55.8% and 56.1% for Bagging, Adaboost and RF respectively. While Bagging takes the lead follow by RF and Adaboost, the level of accuracy obtained dropped significantly compared to when Hungary and Stalog dataset where used.

Therefore, this study will base its recommendation on the result obtained from the dataset used when minimum of 80% accuracy is achieved. The study also highlights the need to improve the considered techniques as errors less than 2% for example might leads to loss of life. Graphically, Figure 8 interpret the same result shown in Table 4a and 4b for easy glance through.

Performance based on other metrics (KS, MCC, ROC and Pr) is presented in Table 4a and Table 4b. These metrics are widely used for evaluation of classification techniques to affirm the authenticity of result derived during experiment.

From Table 4a, the KS metric under dataset with 127 instances shows Bagging takes the lead with 0.16 follow by Adaboost and RF which gives 0.10 and 0.06 respectively. For dataset with 200 instances, the KS obtained shows Bagging also surpasses Adaboost and RF with 0.18, 0.15 and 0.12 respectively. Considering KS metric with 270 instances, 0.66 is obtained for Bagging, 0.64 for Adaboost and 0.62 for RF which also highlight the supremacy of Bagging.

Table 4b includes the KS metrics result where 294 and 303 instances are used for the experiment. Although, there is slight drop in performance of all the algorithms when used on dataset with 294 instances compare to 270 instances but RF show supremacy in terms of KS over Bagging and Adaboost with 0.59, 0.58 and 0.57 respectively. The KS obtained for the techniques when 303 instances of data is used are 0.33, 0.31 and 0.27 for Bagging, Adaboost and RF respectively. Bagging outperformed Adaboost and RF but the general performance is poor compared to the result obtain when 270 and 294 instances of data are used for experiment. This behaviour is similar to the observation made during comparison of the techniques base on accuracy.

For all the experiment scenarios based on number of dataset instances (123, 200, 270, 294 and 303) as shown in Table 4a and Table 4b, the best performance of Bagging, Adaboost and RF in terms of MCC, ROC and Pr is observed when dataset with 270 instances was used for experiment follow by data with 294 instances.

For 270 instances of data in Table 4a, Bagging performs better in MCC with 0.66 compare to 0.64 and 0.62 obtained for Adaboost and Bagging respectively. The ROC and Precision obtain for Bagging and RF in the experiment with 270 instances is the same with 0.90 which is better compare to

Adaboost which gives 0.87 and 0.85 for ROC and Precision respectively.
Considering 294 instances of data in Table 4b, RF takes the lead in terms of MCC with 0.59 follow by Bagging and Adaboost with 0.58 and 0.57 respectively. For ROC and Precision, Bagging takes the lead with 0.89 and 0.88 respectively follow by RF which gives 0.87 and 0.86.

Based on criteria such as partitioning of data into subset, the primary goal, combinatorial function for model and methods where it is used, Table 5 shows the general comprison of Boosting and Bagging technique ([Cro12]).

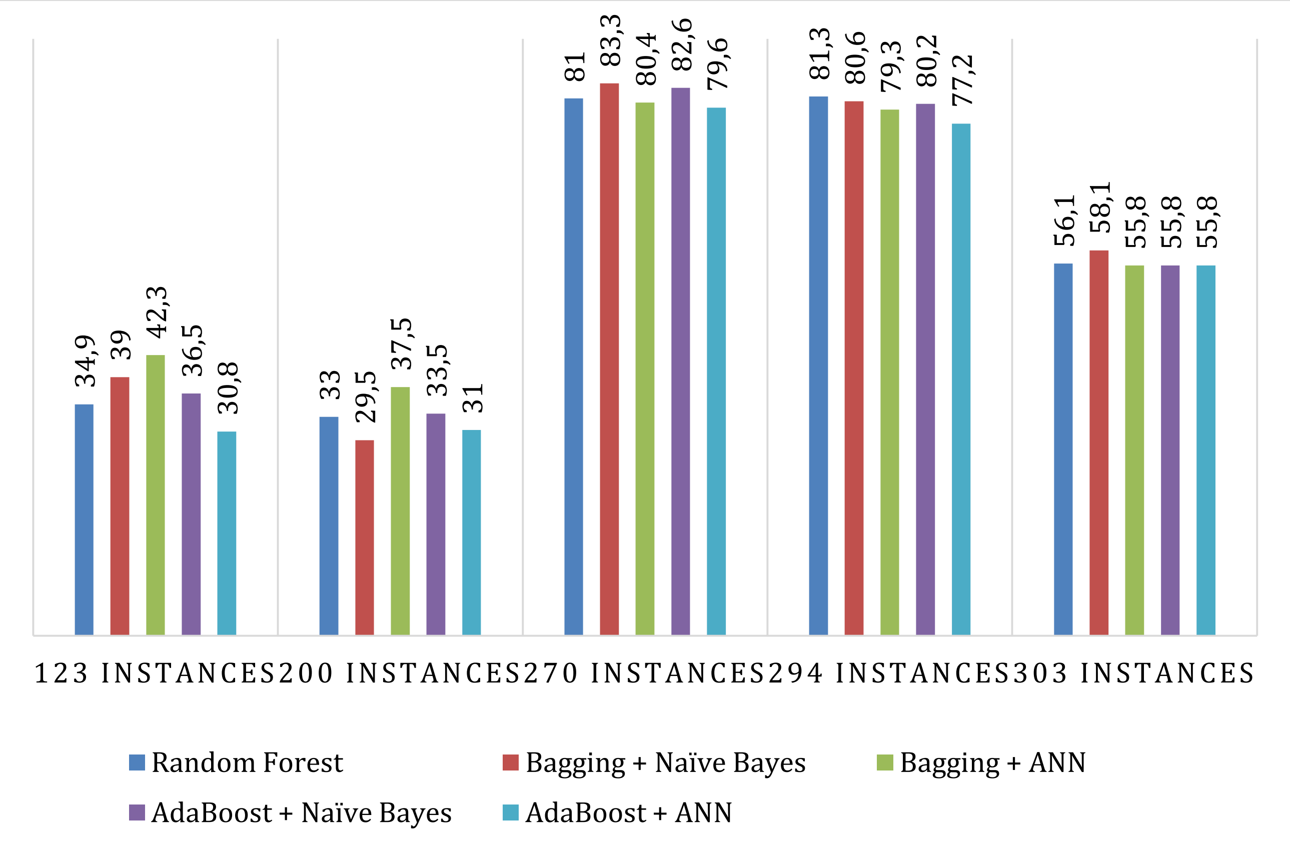

Fig. 8: Graph of the Accuracy Comparison

Table 4a: Comparison base on other parameters

Techniques		123 Instances				200 Instances				270 Instances			
		KS	MCC	ROC	Pr	KS	MCC	ROC	Pr	KS	MCC	ROC	Pr
Random Forest		0.06	0.06	0.58	0.35	0.12	0.12	0.61	0.37	0.62	0.62	0.90	0.90
Bagging	Naïve Bayes	0.11	0.12	0.55	0.32	0.08	0.08	0.61	0.34	0.66	0.66	0.90	0.90
	ANN	0.16	0.15	0.61	0.39	0.18	0.18	0.63	0.35	0.60	0.60	0.88	0.88
AdaBoost	Naïve Bayes	0.10	0.12	0.55	0.34	0.15	0.15	0.61	0.36	0.64	0.64	0.87	0.85
	ANN	0.02	0.02	0.56	0.34	0.10	0.10	0.62	0.37	0.59	0.59	0.86	0.86

Table 4b: Comparison base on other parameters

Techniques		294 Instances				303 Instances			
		KS	MCC	ROC	Pr	KS	MCC	ROC	Pr
Random Forest		0.59	0.59	0.87	0.86	0.27	0.34	0.81	0.60
Bagging	Naïve Bayes	0.58	0.58	0.89	0.88	0.33	0.42	0.81	0.60
	ANN	0.54	0.54	0.86	0.87	0.29	0.37	0.81	0.60
AdaBoost	Naïve Bayes	0.57	0.57	0.86	0.85	0.31	0.39	0.68	0.50
	ANN	0.50	0.50	0.85	0.84	0.30	0.37	0.79	0.56

Table 5: General comparison of Bagging and Boosting Technique based on other criteria

Criteria	Bagging	Boosting
Data Partition into subset	Random	Given mis-classified samples higher prefrence
Primary Goal	Minimize variance	Increase the force for prediction
Methods where this is used	Random subspace	Gradient descent
Function used in combining single models	(weighted) average	Weighted majority vote

5. SUMMARY AND CONCLUSION

The importance of the heart cannot be over emphasised. Because of its functions and how it works, the fact that claims the end of the heart activities in human being is the end of life is unarguable. Therefore, proper care is need for the heart and early treatment in case of any disorder. Meta-algorithms in data mining have been in existence for a long time and are used to solve real life problem almost in all sectors. This paper presents the comparison of two famous ensemble method (Bagging and AdaBoost) that are used to improve the performance of data mining algorithms. Five heart disease datasets with different size are used and the classifiers considered are NB, ANN and RF. The ensembles were only applied to NB and ANN during the experiment while RF is considered as Bagging of DT. The result shows a flaunting performance in terms of accuracy during the experiment. Bagging of NB gives the highest accuracy with 83.3% out of all the collated result.

In terms of other parameters such as Kappa statistic, MCC, ROC Area and Precision, Bagging surpasses AdaBoost in the experiment result. Therefore, study recommends Bagging as a better ensemble method compare to AdaBoost. Although, the best performance was generally observed when dataset of 270 instances was used and there is drop in performance of all the techniques in terms of Accuracy, MCC, KS, ROC and Precision when applied on dataset with larger instances. This raised a high level of concern.

It should be noted that an error rate of 5% or lesser can leads to loss of life during treatment. Therefore, the future work will focus on developing a technique that will be efficient and reliable for prediction of deadly diseases such as the heart disease, which will also be suitable for large dataset and also product acceptable results on mini dataset.

REFERECES

[Bre96] **L. Breiman** – *Bagging predictors.* Machine Learning. 24 (2): 123–140, 1996 CiteSeerX 10.1.1.32.9399 Freely accessible. doi:10.1007/BF00058655.

[Cro12] **V. Cross** – *Bagging, boosting and stacking in machine learning*, 2012. Retrieved from https://stats.stackexchange.com/questions/18891/bagging-boosting-and-stacking-in-machine-learning, 2012 Accessed on 29 December 2017.

[CB15] **T. P. Colville, J. M. Bassert** – *Clinical Anatomy and Physiology for Veterinary Technicians.* Elsevier Health Sciences. p. 547. (10 March 2015). ISBN 978-0-323-35620-6.

[CS12] **S. D. Chaitrali, S. A. Sulabha** – *Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques.* International Journal of Computer Applications (0975 – 888) Volume 47– No.10, pg44–48 June 2012.

[Ens16] **Ensemble Method** – 2016 Retrieved from: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0ahUKEwjXhbyjltDYAhXIC8AKHavxBVIQFggnMAA&url=https%3A%2F%2Fwww3.nd.edu%2F~rjohns15%2Fcse40647.sp14%2Fwww%2Fcontent%2Flectures%2F31%2520-%2520Decision%2520Tree%2520Ensembles.pdf&usg=AOvVaw1bGrTqaXb-9BoqZ0Mr84xp. Accessed on 29 January 2018.

[Jaa16] **T. Jaakkola** – *MIT CSAIL* 2016. Retrieved from: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0ahUKEwj79Przl9DYAhXnJ8AKHeZzA-MQFggnMAA&url=http%3A%2F%2Fpeople.csail.mit.edu%2Fdsontag%2Fcourses%2Fml12%2Fslides%2Flecture13.pdf&usg=AOvVaw08f-GlrXx3f8QGDEzqtvxA.

[JTS16] **P. Jaymin, U. Tejal, P. Samir** – *Heart Disease Prediction Using Machine learning and Data Mining Technique*, 2016. Vol.7, No.1, pg129–137, Sept 2015-March 2016. DOI: 10.0905992/IJCSC.2016.018. www.csjournalss.com.

[Kee17] **T. K Keerthana** – *Heart Disease Prediction System using Data Mining Method.* International Journal of Engineering Trends and Technology (IJETT) – Volume 47 Number 6 May 2017 Page 361. ISSN: 2231-5381 http://www.ijettjournal.org

[K+14] **P. Kolh, S. Windecker, F. Alfonso, J. Collet, J. Cremer, V. Falk, G. Filippatos, C. Hamm, S. Head** – *2014 ESC/EACTS Guidelines on myocardial revascularization: The Task Force on Myocardial Revascularization of the European Society of Cardiology (ESC) and the European Association for Cardio-Thoracic Surgery (EACTS). Developed with the special contribution of the European Association of Percutaneous Cardiovascular Interventions (EAPCI).* European Journal of Cardio-Thoracic Surgery: Official Journal of the European Association for Cardio-Thoracic Surgery, 2014. 46 (4): 517–592. (October 2014) doi:10.1093/ejcts/ezu366. ISSN 1873-734X. PMID 25173601.

[LJ16] **A. Londhe, A. Joglekar** – *Analysing the risk of heart disease using ANN.* International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 5 Issue 3, March 2016 569 ISSN: 2278 – 1323.

[L+17] **X. Liu, X. Wang, Q. Su, M. Zhang, Y. Zhu, Q. Wang** – *A Hybrid Classification System for Heart Disease Diagnosis Based on the RFRS Method, 2017.* Hindawi Computational and Mathematical Methods in Medicine Volume 2017, Article ID 8272091, 11 pages https://doi.org/10.1155/2017/8272091.

[N+16] **A. Narassiguin, M. Bibimoune, H. Elghazel, A. Aussem** – *An extensive empirical comparison of ensemble learning methods for binary classification, Pattern Analysis and Applications*, 2016 https://doi.org/10.1007/s10044-016-0553-z.

[P+16] **P. Ponikowski, A. A. Voors, S. D. Anker, H. Bueno, J. G. F. Cleland, A. J. S. Coats, V. Falk, J. R. González-Juanatey, V. Harjola** – *2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: The Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC). Developed with the special contribution of the Heart Failure Association (HFA) of the ESC".* European Journal of Heart Failure. 18 (8): 891–975. (August 2016). doi:10.1002/ejhf.592. ISSN 1879-0844. PMID 27207191.

[RPJ14] **C. P. Reddy, P. Palagi, Jaya** – *Heart Disease Prediction Using Ann Algorithm in Data Mining*. International Journal of Computer Science & Mobile Computing, 2017. Vol.6 Issue.4, April- 2017, pg. 168-172 © 2017, IJCSMC Available Online at www.ijcsmc.com.

[RP16] **K. Ramandeep, K. Prabhsharn** – *A Review - Heart Disease Forecasting Pattern using Various Data Mining Techniques*. International Journal of Computer Science and Mobile Computing, IJCSMC, 2016 Vol. 5, Issue. 6, pg. 350 – 354 ISSN 2320–088X Available Online at www.ijcsmc.com.

[Sch13] **R. E. Schapire** – *Explaining adaboost In Empirical Inference*, page 1-52. Springer, 2013. Available at https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0ahUKEwiEjK3D9s_YAhWFxxQKHW0SB4MQFggnMAA&url=http%3A%2F%2Frob.schapire.net%2Fpapers%2Fexplaining-adaboost.pdf&usg=AOvVaw1yiONHF3GQRdgcP5uHwRvw.

[Sin10] **A. Singh** – *Boosting Can We Make Dump learners smart?* October 2010. Accessed: November 2016. Retrieved from: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=

rja&uact=8&ved=0ahUKEwjC2MDwmNDYAhUMKMAKHRP3BqkQFggnMAA&url=https%3A%2F%2Fwww.cs.cmu.edu%2F~aarti%2FClass%2F10701%2Fslides%2FLecture10.pdf&usg=AOvVaw3uH_hSjEMusdd3M8bBvRR6.

[SR17] **A. K. Shrivas, K. Y. Rajat** – *An Effective Prediction Factors for Coronary Heart Disease using Data Mining based Classification Technique.* International Journal on Recent and Innovation Trends in Computing and Communication, 2017. Volume: 5 Issue: 5 813 – 816 ISSN: 2321-8169.

[SS14] **K. Saravanan, S. Sasithra** – *Review on Classification Based on Artificial*. International Journal of Ambient Systems and Applications (IJASA), 2014. Vol.2, No.4, pg11–18, December 2014. https://doi.org/10.5121/ijasa.2014.2402.

[SS16] **C. Sowmiya, P. Sumitra** – *Comparative Study of Predicting Heart Disease by Means of Data Mining*. International Journal of Engineering and Computer Science (IJECS), 2016. Volume 5(Issue 12), 19580-19582. Retrieved from www.ijecs.in.

[VB17] **C. Vincy, M. S. Bindu** – *Heart Disease Prediction Using Naïve Bayes Algorithm and Laplace Smoothing Technique*, 2017. International Journal of Computer Science Trends and Technology (IJCST) – Volume 5 Issue 2, Mar – Apr 2017 ISSN: 2347-8578 www.ijcstjournal.org.

[VS13] **C. Vikas, P. Saurabh** – *Prediction of Heart Diseases Using Data Mining Techniques*. Caribbean Journal of Science and Technology, Vol.1, 208-217, 2013. ISSN 0799-3757, http://caribjscitech.com/.

[V+12] **A. Vahanian, O. Alfieri, F. Andreotti, M. Antunes, G. Barón-Esquivias, H. Baumgartner, M. A. Borger, T. P. Carrel, D. B. Michele** – *Guidelines on the management of valvular heart disease (version 2012): The Joint Task Force on the Management of Valvular Heart Disease of the European Society of Cardiology (ESC) and the European Association for Cardio-Thoracic Surgery (EACTS)*. European Journal of Cardio-Thoracic Surgery: Official Journal of the European Association for Cardio-Thoracic Surgery. 42 (4): S1–44. (October 2012). doi:10.1093/ejcts/ezs455. ISSN 1873-734X. PMID 22922698.