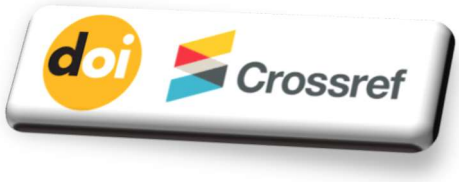
Journal, Advances in Mathematical & Computational Sciences
Vol. 10 No. 3, 2022
www.isteams.net/mathematics-computationaljournal



Journal of Advances in Mathematical & Computational Sciences
An International Pan-African Multidisciplinary Journal of the SMART Research Group
International Centre for IT & Development (ICITD) USA

Available online at https://www.isteams.net/ mathematics-computationaljournal.info
CrossREF Member Listing - https://www.crossref.org/06members/50go-live.html

A Mobile Adaptive Online Diagnostics Machine Learning System for Heart Disease Prediction

[1]Ezea, Ikenna L. & [2]Uba, Janefrances Ifeyinwa
[1]Department of Computer Science/Informatics, Alex Ekwueme Federal University, Ndufu-Alike, Ebonyi State, Nigeria
[2]Department of Computer Engineering Department, Cyprus International University Nicosia, Northern Cyprus, Via Mersin 10, Turkey
E-mails: ezeaikenna@yahoo.com; uba.janefrances@gmail.com

ABSTRACT

Heart Disease prediction has become a trending topic in the Machine Learning Community due to the prevalence of the disease and the high mortality rate. Most complications associated with this diseases are due to delay in diagnoses which in effect affect early treatment. The number of casualties could be reduced if the latency of prediction and treatment is reduced. This research addressed this issue by developing a mobile adaptive online diagnostic system that helps in early diagnosis of this disease. To achieve this we benchmarked different machine learning algorithms and picked Random Forest which had the best performance among all the tested algorithms. To improve performance Entropy and Information Gain feature extraction technique was used. The result of this study showed that heart disease diagnosis could be done across different devices, both computer and mobile. This development will make the accessibility of this services to the remote users who may not have computer systems but can afford mobile phones.

Keywords: Machine Learning, Algorithms, Applications, Predictions, Deductions, Analytics

Eaea, I.L. & Uba, J.I. (2022): A Mobile Adaptive Online Diagnostics Machine Learning System for Heart Disease Prediction. Journal of Advances in Mathematical & Computational Science. Vol. 10, No. 3. Pp 15-28
DOI: dx.doi.org/10.22624/AIMS/MATHS/V10N2P2.
Available online at www.isteams.net/mathematics-computationaljournal.

1. MACHINE LEARNING

Medical Diagnosis is an important medical procedure that should be initiated in other to address any medical problem. It is the process of identifying a disease based on the symptoms. Depending on the disease the process could be simple or complicated and time consuming.

Heart Diseases which is a disease that affects the heart or the blood vessels and consequently prevents them from carrying out their major function is one type of disease that takes a long process to diagnose. A heart disease could be reversed or managed if the diagnoses is timely. Consequently, most research in machine learning have been channeled towards having efficient, timely and convenient heart disease diagnostic systems. The idea has been to make predictions using features that are easily accessible from patients rather than subjecting them to rigorous and expensive medical diagnostic procedures. Several Machine Learning algorithms have been considered for classification [1], prediction [2], and diagnoses [3]. Most of these algorithms were to contend with making tradeoffs between accuracy and algorithm complexity.

However, the contending issue is not just accuracy and simplicity of the algorithm but delivery of solution that will balance the tradeoff to deliver a mobile adaptive solution. Most people due to some reasons bothering on cost, convenience, power supply, and work demands may not opt for personal computers, however they cannot do without mobile phones. Thus the need for mobile and adaptive system which will not just deliver a convenient solution but also deliver a solution that will be accessible to both personal computer (pc) and mobile phone users.

This ensures that the solution will be accessible to all, both the urban dwellers who have the highest population of users with both PC's and mobile phones and the rural dwellers who may be dominated by users with just mobile phones. Thus this work aims to build a mobile and adaptive system for the diagnosis of hearth diseases. The algorithm used for the system implementation is Random Forest which had the best score after testing other five major machine learning algorithms (Support Vector Machine, Naïve Bayes, Logistic Regression, K-Nearest Neighbour and Decision Tree Classifier) on an optimal dataset features. After the implementation one is expected to have a system that can diagnose heart disease across mobile phone platforms with an accuracy of 100%.

2. LITERATURE REVIEW

Machine Learning research has revolutionized Heart Disease (HD) prediction by providing different flavors of algorithms. Most of the algorithms have recorded great success with varying degrees of accuracies. Depending on application domain there may be some tradeoff between the algorithm accuracy and computational demands (time and complexity). While some algorithms produced result with high level of accuracies they may be computationally intensive. This has been the major drawback in most research in heart disease predictions algorithms.

Gudadhe et al. [4] combined two machine learning algorithms (Multilayer Perceptron and Support Vector Machine) for Heart Disease classification. This approach gave a performance accuracy of 80.41%. Even though the performance is relatively high the approach is computational intensive. The same also is applicable to Resul et al. [5] who developed Heart Disease ensemble classification system using Artificial Neural Network with a performance accuracy of 89.01%, a sensitivity rate of 80.09% and specificity of 95.91%. Another scholars that had a similar complexity tradeoff are Samuel et al. [6] that achieved performance accuracy of 91.10% using Decision Support System and Fuzzy AHP for Heart Disease diagnoses. Additionally, some scholars have recorded relatively high performance accuracy with high computational time.

That is a tradeoff between accuracy and computational time. Olaniyi et al. [7] used three phase Artificial Neural Network Technique to predict Heart Disease in Angina. Though the performance accuracy was 88.89% the computational time was high. The same also goes to Liu et al. [8] who used Relief and Rough technique for Heart Disease classification at a performance accuracy of 92.32%. Furthermore, some of the algorithms that achieved low computational time had relatively low performance accuracy. MOHAN et al. [9] performed Heart Disease prediction using Hybrid Machine Learning technique. They achieved a low computational time with an accuracy of 88.07%. Derano et al. [10] on the other hand developed a heart disease classification system using Machine Learning algorithm. They used Cleveland dataset and achieved a performance accuracy of 77%. Their approach was computationally less complex.

Depending on the computational resources available and volume of training data one can make a tradeoff that favours one (performance, complexity and time) against the other or rather balance between the three main tradeoffs. However, every consideration should favor performance accuracy since human health is involved. This is the reason why we used Random Forest algorithm which gave a performance accuracy of 100% with manageable computational demands.

Table 1: Performance of different reviewed methods

Author	Objective	Technique	Accuracy
Detrano et al. [10]	HD classification	Multilayer classification Technique	77.00%
Gudadhe et al. [4]	HD Classification	MLP and SVM	80.41%
Kahramanli et al. [11]	HD Classification	NN and Fuzzy Logic	87.40%
Das et al. [5]	HD Classification	ANN and SMEM	89.01%
Palaniappan et al. [12]	HD Identification	Expert System + NB, DT and ANN	NB (86.12%), DT (80.41%), ANN (88.12%)
Olaniyi et al. [7]	HD Prediction	Three phase ANN	88.89%
Samuel et al. [6]	HD Diagnosis	DSS and Fuzzy AHP	91.10%
Liu et al. [8]	HD Classification	Relief and Rough	92.32%
Mohan et al. [9]	HD Prediction	Hybrid ML + Feature Selection	88.07%

3. MATERIALS AND METHODS

3.1 Dataset Description

The dataset used in this work was gotten from Kaggle dataset repository. It consists of 1025 records and 14 features. Fourteen (14) out of the 1025 records were removed due to missing values. The dataset consists of two categories of samples that make up the target feature. The first category which is 0 (heart disease absent) has 499 records while the second category which is 1 (heart disease present) has 512 records. This brings the total records to a new value of 1011. This means the dataset is fairly balanced. The attributes of the datasets are shown in table 2.

Table 1: Dataset Attributes and Description

SNO	Attribute	Attribute Code	Description
1	Age	AGE	Age in years
2	Sex	SEX	Gender (1 = male; 0 = female)
3	Chest Pain	CPT	chest pain type
4	Resting Blood Pressure	RBP	resting blood pressure (in mm Hg on admission to the hospital)
5	Serum Cholesterol	CHL	serum cholesterol in mg/dl
6	Fasting Blood Sugar	FBS	(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
7	Resting Electrocardiographic Result	ECG	resting electrocardiographic results
8	Exercise Induced Angina	EIA	exercise induced angina (1 = yes; 0 = no)
9	Sinus Tachyeardia Depression Induced by Exercise Relative to Rest	OPK	ST depression induced by exercise relative to rest
10	Slope of the Peak Exercise ST Segment	SLP	the slope of the peak exercise ST segment
11	Number of Major Vessels	NMV	number of major vessels (0-3) colored by fluoroscopy
12	Thallium Stress Test Result	THL	3 = normal; 6 = fixed defect; 7 = reversible defect
13	Expected Output	OUT	have disease or not (1 = yes; 0 = no)

Table 2: Model Execution Time and Accuracy

Model	Test Accuracy	Execution Time (s)
Logistic Regression (LR)	85.85%	0.032201052
K-Nearest Neighbour (KNN)	100.00%	0.030536175
Support Vector Machine (SVM)	91.71%	0.035356522
Naïve Bayes Classifier (NB)	88.29%	0.032014847
Decision Tree (DT)	100.00%	0.010361195
Random Forest (RF)	100.00%	2.109589100

3.2. System Specification

The implementation of this project was done using hardware and software tools. A Hewlett Packard (HP) computer system and a couple of IDE's and application development libraries were used for the code development. The details of the exact hardware and software specification is shown in table 4 and table 5 respectively.

Table 3: Computer System Specification

Compoent	Specification
Operating System	Windows 10 Pro
Processor	Intel ® Pentium CPU N3710 @ 1.60GHz
Installed Memory (RAM)	4.00 GB
System Type	64-bit Operating System

Table 4: Application Development Tools

Component	Specification
Python IDE	Google Colab
Python Library	Pandas, Sklearn, Matplotlib, Numpy, Seaborn
Database System	Mysql 8.0
File System	MS Excel CSV
Web Development Tool	HTML, CSS, Javascript
Java IDE	Eclipse

3.3. The Proposed System Framework.

Figure 1 shows the shows the series of activities undertaken to get the desired result. The phases involved in the framework are:

3.3.1. Dataset Cleaning

The state of the dataset determines the performance of any model. A dataset in an inconsistent state (too many redundant and missing values) is likely going to perform poorly. So at this stage the dataset was cleaned so as to get it in a consistent state.

3.3.2. Feature Selection

Some features in the dataset are redundant and may not impact the classification performance. Thus training with such features will affect the performance of the system. So on this note we used Spearman Correlation Coefficient for the dataset features. Feature selection is also important in determining the decision node in a decision tree in that case we used entropy and information gain (see equation 7-9). The Spearman Correlation expression is as follows:

3.3.3. Dataset Splitting

This is the process of dividing the dataset into two unequal parts which will be used for training and testing. This ensures that bias and variance issues are controlled. The dataset used for this experiment was split in the ratio of 80:20 that is 80% was used for training while 20% was used for testing.

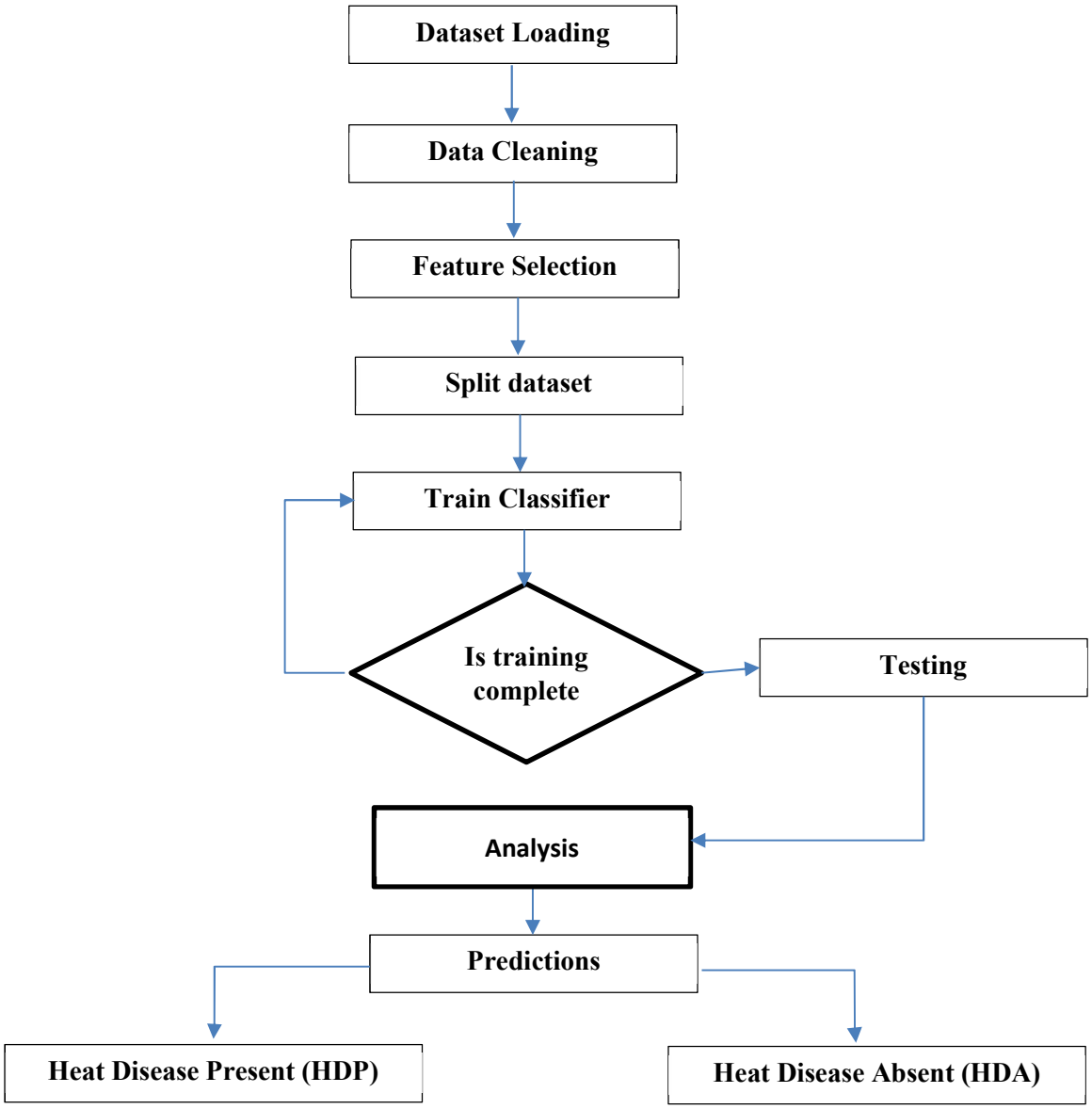


Figure 1: Proposed Heart Disease Classification System Framework

3.3.4. Training Classifier. The goal is to adjust the parameters of the model until the targets are well separated from each other. This helps in making accurate prediction whenever a data sample is provided.

3.3.5. Testing
This involves using the test data on the model to know the distribution of errors. The performance of the model is evaluated at this point using accuracy and confusion matrix (see table 3 and table 7 respectively).

3.3.6. Analysis
This involves the evaluation of the model based on the following performance metrics: Receiver Operator Characteristics (ROC), Confusion Matrix and Accuracy. The choice of a given classifier depends on how it competes with others using the above performance metrics.

3.3.7. Prediction
At this stage the system is expected to determine if a given data sample has heart disease or not. If predicts the presence of heart disease and as well give the percentage chances of developing heart disease.

3.4. Comparative Algorithms
An algorithm is the sequence of steps to be followed in carrying out a given task. In Machine Learning there are many algorithms that can perform a similar task like classification, regression, etc. They all have the ability of given the same or similar result but with different level of accuracy or computational time. The choice of algorithm and performance tradeoff to be made in any Machine Learning project depends on available equipment and subject domain. Being that this research is on health the priority must be on accuracy. So this section presents all the Machine Learning algorithms to be evaluated for the experiment.

3.4.1. Logistic Regression (LR) is a supervised machine learning algorithm that classifies data using the concept of probability. This can be shown using the following mathematical expression [11]:

$$P(X) = \frac{\exp(\beta_0 + \beta_1 x)}{1+\exp(\beta_0 + \beta_1 x)} \qquad (1)$$

3.4.2. K-Nearest Neighbour is a non-parametric supervised machine learning algorithm that splits a datasets into related groups based on the proximity of the data points. This algorithm can be used for classification as well as regression and it uses the concepts of different distance formula such as Manhattan, Minkowski, Hamming Distance and Euclidean Distance which was used in this article and it is expressed as follows:

$$d(X,Y) = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2} \quad (2)$$

3.4.3. Support Vector Machine is supervised learning models that consists of algorithms that are used for analysis of data for classification and regression. The support vector machine can be expressed using the following equations [12]:

$$H_0 : \mathbf{w^T x} + b = 0 \qquad (3)$$

$$H_1 : \mathbf{w^T x} + b = -1 \qquad (4)$$

$$H_2 : \mathbf{w^T x} + b = 1 \qquad (5)$$

Where H_1 and H_2 are the planes and H_0 is the median in between the planes, **w** is the weight vector, **x** is the input vector and b is the bias.

3.4.4. Naïve Bayes Algorithm is a probabilistic Machine Learning classification algorithm that is based on the following Bayes Theorem.

$$P(A|B) = \frac{P(B|A)*P(A)}{P(B)} \qquad (6)$$

Where P(A|B) is the probability of A occurring given that B has already occurred,

P(B) is the probability of B, P(A) is the probability of A and P(B|A) is the probability of B occurring given that A has already occurred.

3.4.5. Decision Tree

This Is a supervised machine learning algorithm that uses the concept of entropy and information gain to split the dataset for classification and regression. The entropy (H(S)) is used to calculate the amount of randomness in data while the information gain (IG(S, A)) is the amount of information to be gained if a given attribute is taken as the node (decision point) for the split of the data. The algorithmic concepts is as shown below:

$$H(S) = \sum_{x \in X} P(x) \ln \frac{1}{P(x)} \qquad (7)$$

$$IG(S, A) = H(S) - H(S, A) \qquad (8)$$

Alternatively

$$IG(S, A) = H(S) - \sum_{x \in X} P(x) * H(x) \qquad (9)$$

Where P(x) is the probability of event x, H(S, A) is the effective change in entropy after the attribute A is chosen and H(x) is the entropy of x.

3.4.6. Random Forest

This is a supervised machine learning algorithm that uses bootstrapping and aggregation for classification and regression of data samples. It is an extension of decision tree with the inclusion of multiple random decision trees. In comparison with decision tree it has a minimal overfitting and higher performance accuracy [13]. The pseudocode can be shown in the figure 2.

Precondition: A training set S := (x_1, y_i), …, (x_n, y_n), features F, and number of trees in forest B

```
1.  Function RANDOMFOREST(S, F)
2.      H ← 0
3.      for i ϵ 1, …., B do
4.          S(i) ← A bootstrap sample from S
5.          hi ← RANDOMIZEDTREELEARN(S(i), F)
6.          H ← H U {hi}
7.      end for
8.      return H
9.  end function
10. function RANDOMIZEDTREELEARNS(S, F)
11.     At each node:
12.         f ←very small subset of F
13.         split on best feature in f
14.     Return The learned tree
15. end function
```

Figure 2: Random Forest Algorithm

3.5. Model Evaluation and Selection

The evaluation of the model helps to access the model based on their performance of which the best is selected for the implementation. The following metrics were considered in the evaluation of the algorithms used for this work.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+F}} \qquad (10)$$

Where TP is True Positive, TN is True Negative, FP is False Positive and FN is False Negative.

3.6. Spearman Rank Correlation Coefficient

This is a measure of the strength of the difference between different features in the dataset. This was used in this research to determine which feature to be included based on the strength of the relationship with other features in the dataset. The expression is as follows:

$$r = 1 - \frac{6\sum d^2}{n^3 - n} \qquad (11)$$

Where r is the correlation coefficient, d is the difference between the ranks and n is the number of records in the dataset.

3.5.1. ROC (Receiver Operator Characteristics)

Depicts the performance of a classification model at every given classification threshold. It consists of two parameters (True Positive and False Positive Rate) at the y-axis and x-axis respectively: The ROC curve shown in figure 3 shows an Area under the Curve ROC (AUROS) with result of 1.000 for RF, KNN and DT.

- True Positive Rate (at the y-axis)
- False Positive Rate (at the x-axis)

True Positive Rate (TPR) is also called the recall and it can be expressed in the following formula.

$$\text{TPR (Recall)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad (12)$$

False Positive Rate (FPR) can be expressed in the following formula:

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \qquad (13)$$

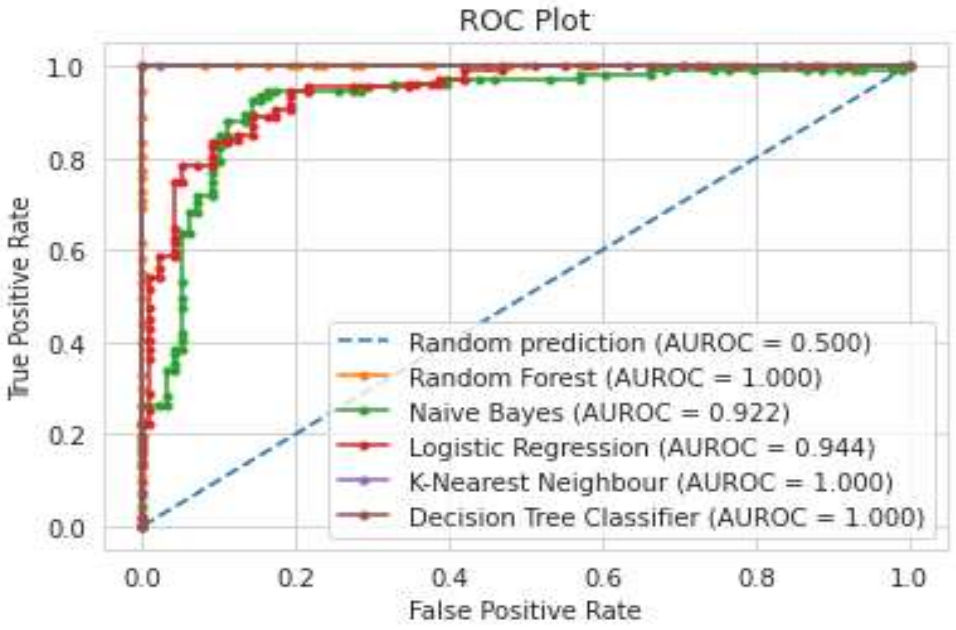


Figure 3: Receiver Operator Characteristics Curve

3.5.2. Confusion Matrix. The confusion matrix shows the distribution of errors with regards to some evaluation parameters (see table 6) associated with the algorithm models. The most valuable parameters are the diagonal values which shows how many predicted values were actually positive (TP) and negative (TN).

The smaller the number in the FP and FN the better as these are the errors in prediction. Table 7 shows the actual result gotten from the six algorithms considered for this paper and it shows that Random Forest (RF) had the best performance with TP = 98, TN = 107, FP = 0 and FN = 0.

Table 5: The Template used for the Confusion Matrix in Table 7

		Predicted Values	
		Positive	Negative
Actual Value	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Table 6: Confusion Matrix for the six Machine Learning Algorithms

Algorithm	Confusion Metrix	
Logistic Regression (LR)	79	19
	10	97
K-Nearest Neighbour (KNN)	96	2
	0	107
Support Vector Machine (SVM)	91	7
	10	97
Naïve Bayes (NB)	84	14
	10	97
Decision Tree (DT)	84	14
	10	97
Random Forest (RF)	98	0
	0	107

4. RESULT AND DISCUSSION

Different experiments was conducted to ensure that the right procedures was followed to arrive at the desired result. The first phase of the experiment was to benchmark and select the best among the six most popular machine learning algorithms. The result as can be seen in table 3 shows that Random Forest, Decision Tree and K-Nearest Neighbour with accuracy of 100% outperformed the rest of the algorithms.

Meanwhile, the result in table 7 shows that RF outperformed DT and KNN in the number of False Positive and False Negatives. However RF has the highest computational demand (see table 3) and that would have given DT an edge if not for the fact that DT has the tendency of overfitting due to its low bias and high variance nature. Consequently, the experimental consideration that have led to the choice of RF as the ideal algorithm for the implementation laid the ground work for the mobile adaptive system which is not just very efficient but has a high response rate. The data input interface for the application is shown in figure 7 while the mobile phone and pc view of the heart disease prediction is shown in figure 5 and figure 4 respectively. The results shows the patients input and the percentage progression in the patient's chances of developing heart disease.

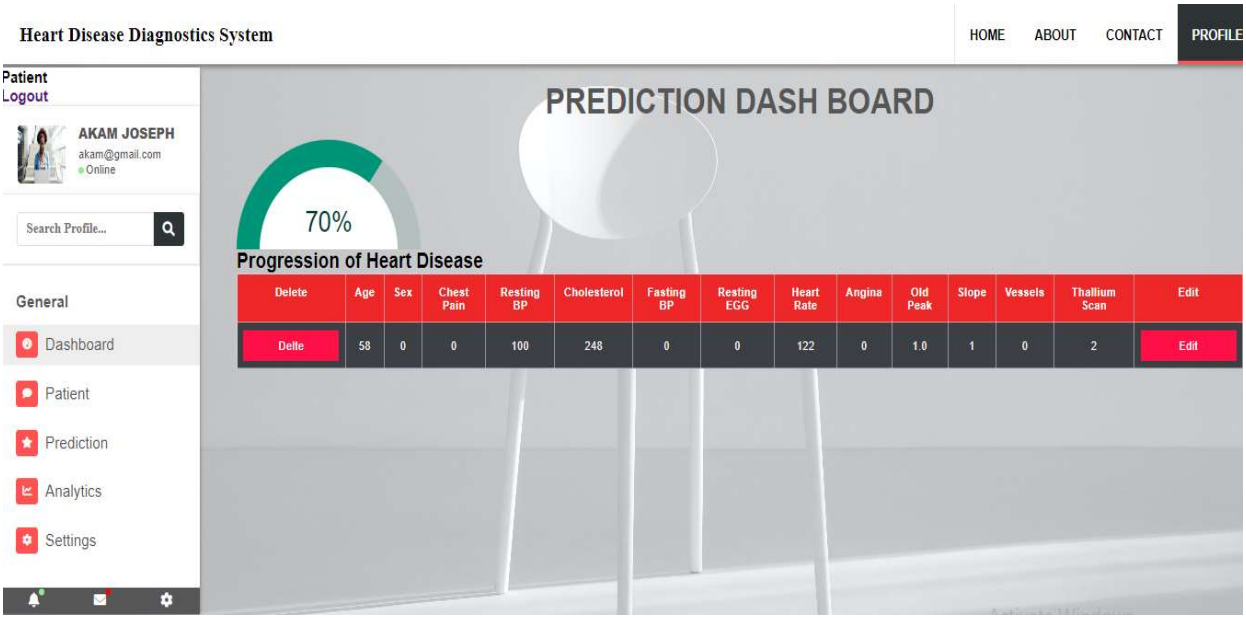


Figure 4: A Personal Computer (PC) Dashboard For Heart Disease Prediction

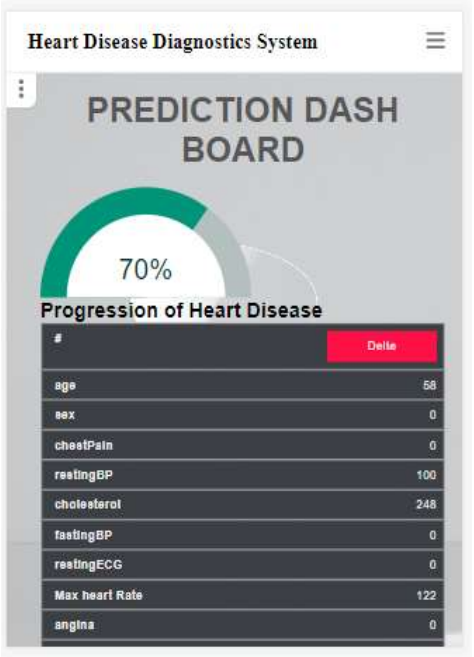


Figure 5: A Mobile Phone View Of A Patient's Heart Disease Diagnosis

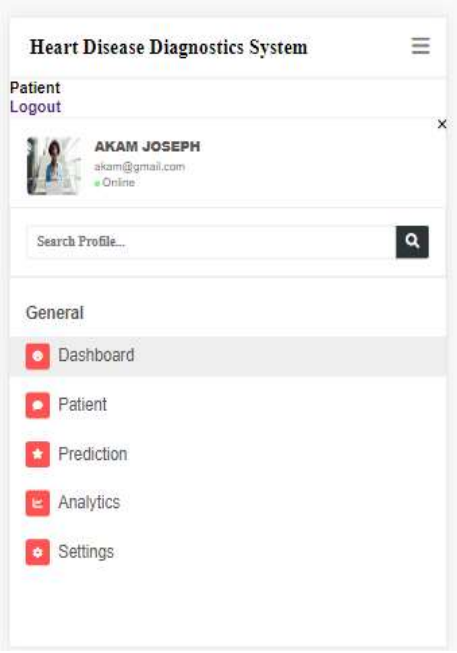


Figure 6: A Mobile Phone navigation link

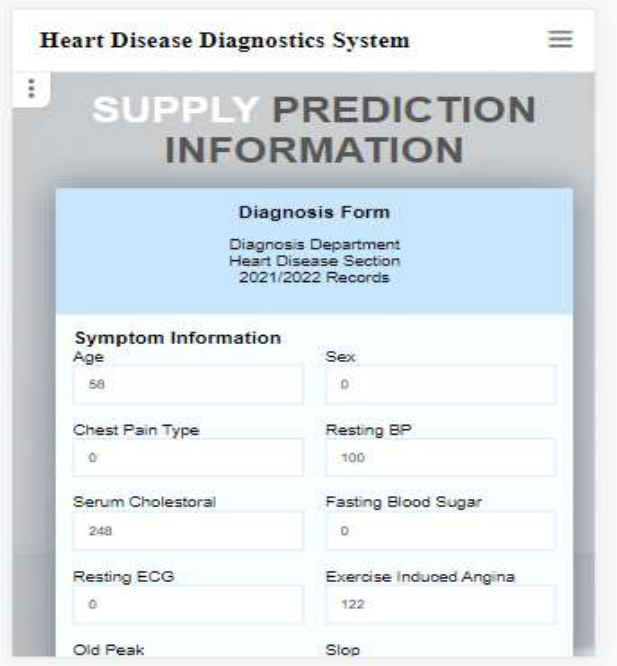


Figure 7: A Mobile Phone View of Patient's Diagnosis form

5. CONCLUSION

Having a mobile responsive heart disease diagnostics system implies that greater percentage of patients will be able to access the diagnostic services. This by implication will reduce the mortality and spread of heart disease as early diagnosis helps prevent complication and make reversal and management possible.

REFERENCES

[1] M. A. Jabbar, B. Deekshatulu and P. Chandra, "Classification of heart disease using artificial neural network and feature subset selection,," Global J. Comput. Sci. Technol. Neural Artif. Intell, vol. 13, no. 3, pp. 4-8, 2013.

[2] A. U. Haq, J. Li, M. H. Memon, M. H. Memon, J. Khan and S. M. Marium, "Heart disease prediction system using model of machine learning and sequential backward selection algorithm for features selection," in IEEE 5th Int. Conf. Converg. Technol. (ICT), 2019.

[3] S. Ghwanmeh, A. Mohammad and A. Al-Ibrahim, "Innovative artificial neural networks-based decision support system for heart diseases diagnosis," J. Intell. Learn. Syst. Appl., vol. 5, 2013.

[4] M. Gudadhe, K. Wankhade and S. Dongre, "Decision support system for heart disease based on support vector machine and artificial neural network," in Int. Conf. Comput. Commun. Technol. (ICCCT), 2010.

[5] R. Das, I. Turkoglu and A. Sengur, "Effective diagnosis of heart disease through neural networks ensembles," Expert Syst Appl, vol. 36, no. 4, pp. 7675-7680, May 2009..

[6] O. W. Samuel, G. M. Asogbon, A. K. Sangaiah, P. Fang and G. Li, "An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction," Expert Syst. Appl, vol. 68, pp. 163-172, Feb 2017.

[7] E. O. Olaniyi, O. K. Oyedotun and K. Adnan, "Heart diseases diagnosis using neural networks arbitration," Int. J. Intell. Syst. Appl, vol. 7, no. 12, p. 72, 2015.

[8] X. Liu, X. Wang, Q. Su, M. Zhang, Y. Zhu, Q. Wang and Q. Wang, "A hybrid classification system for heart disease diagnosis based on the RFRS method," Comput. Math. Methods Med, vol. 2017, pp. 1-11, Jan 2017.

[9] S. Mohan, C. Thirumalai and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," IEEE Access, vol. 7, pp. 81542-81554, 2019.

[10] R. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, J.-J. Schmid, S. Sandhu, K. H. Guppy, S. Lee and V. Froelicher, "International application of a new probability algorithm for the diagnosis of coronary artery disease," Amer. J. Cardiol, vol. 64, no. 5, pp. 304-310, Aug 1989.

[11] H. Kahramanli and N. Allahverdi, "Design of a hybrid system for the diabetes and heart diseases," Expert Syst. Appl., vol. 35, no. 1-2, pp. 82-89, Jul. 2008.

[12] S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," in IEEE/ACS Int. Conf. Comput. Syst. Appl, Mar. 2008.

[13] E. Boateng and D. Abaye, "A Review of the Logistic Regression Model with Emphasis on Medical Research," Journal of Data Analysis and Information Processing, pp. 190-207, 2019.

[14] D. A. Pisner and D. M. Schnyer, "Chapter 6 - Support vector machine," in Machine Learning, , ISBN 9780128157398,, Academic Press, 2020, pp. 101-121.

[15] A. Paul, D. P. Mukherjee, P. Das, A. Gangopadhyay, A. R. Chintha and S. Kundu, "Improved Random Forest for Classification," in IEEE Transactions on Image Processing, 2018.