Full Paper

PREDICTIVE MODELS FOR HEART ATTACK DISEASE RISK

I. O. Awoyelu
Department of Computer Science and Engineering, Obafemi Awolowo University, Ile-Ife, Nigeria
iawoyelu@gmail.com

Y. Egbekunle
Department of Computer Science and Engineering, Obafemi Awolowo University, Ile-Ife, Nigeria

O. Ogunlade
Department of Physiological Sciences, College of Health Sciences, Obafemi Awolowo University, Ile-Ife, Nigeria

ABSTRACT

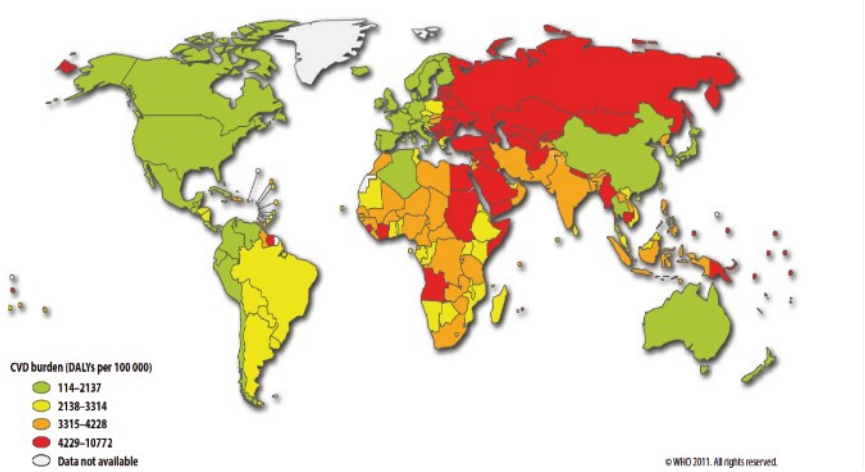
Data mining techniques have been widely applied to the diagnosis of heart diseases. This paper develops models that predict the risk of an individual developing a heart attack. This is with a view to formulating a predictive model to determine the likelihood of having heart attack in a patient, simulating the model formulated, and evaluating the performance of the model. The predictive models were formulated and simulated using Rapid Miner Studio - a collection of knowledge discovery and machine learning tools. The models were developed using Decision-tree, Naïve-Bayes, and Bagging classifiers. The inputs for these prediction models were risk factors of heart attack. A hybrid feature selection technique was carried out and the features selected were used as input variables to the classifiers. 10-fold cross validation was used to assess the performance of the algorithms in predicting a class. The models' performance was evaluated using accuracy and sensitivity as metrics. The Naïve Bayes model exhibits a better performance with an accuracy of 87.86%. The model is expected to enhance the decision-making process of cardiologist.

Keywords: Decision-Tree, Bagging, Naïve-Bayes, Feature-Selection, Information-Gain-Ratio, Wrapper, Heart-attack.

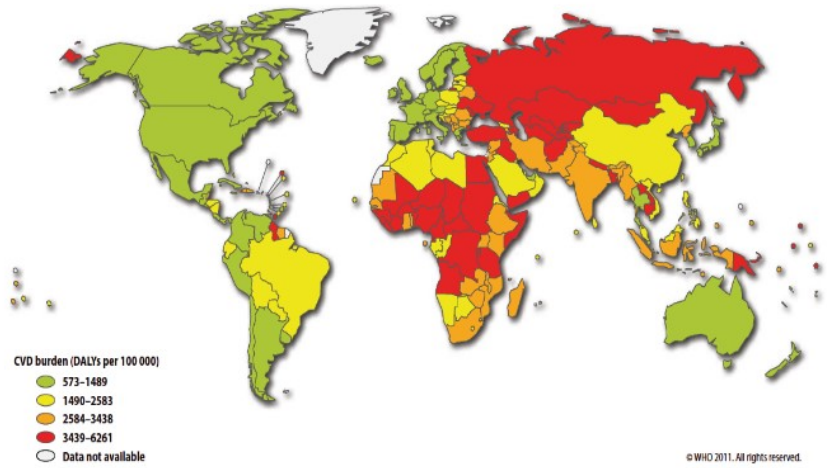
1. INTRODUCTION

Cardiovascular disease includes a wide range of conditions that affect the heart, blood vessels and the way blood is pumped and circulated through the body (Soni *et al*., 2011). Cardiovascular disease (CVD) is responsible for a large proportion of deaths and disabilities worldwide. However, a substantial portion of the increasing global impact of CVD is attributable to economic, social, and cultural changes that have led to increases in risk factors for CVD (Oguanobi *et al*., 2013). These changes are most pronounced in the countries comprising the developing world. Because most of the world's population lives in the developing world, the increasing rate of CVD in these countries is the driving force behind the continuing dramatic worldwide increase in CVD. Until recently, the burden of non-communicable diseases (NCDs) was thought to be a problem afflicting only affluent countries. However, emerging evidence has indicated that the problem affects the developing nations more than the developed ones (World Health Organization, 2008). Figure 1 shows the global CVD mortality(death) rates in males and females, in which Nigeria falls in the category of three hundred and sixty-three thousands to four hundred and forty three thousand. Figure 2 shows the global disease burden due to CVDs in males and females, (WHO, 2011). The disability-adjusted life year (DALY) is a measure of overall disease burden, expressed as the number of years lost due to ill-health, disability, or early death (premature death). Hence CVDs and their risk factors are major contributors to global morbidity and mortality (WHO, 2007; WHO, 2009). Cardiovascular diseases (CVD) result in several illnesses, disabilities, and deaths (WHO, 2007; WHO, 2009; Roger et al., 2012; Mackay and Mensa, 2004). The most common causes of heart diseases are atherosclerosis and/or hypertension (Joshi and Nair, 2015).

Atherosclerosis is a condition that develops when a substance called plaque builds up in the walls of the arteries. This build-up narrows the arteries, making it harder for blood to flow through. If a blood clot forms, it can stop the blood flow. This can cause a heart attack (Joshi and Nair, 2015). Heart attack occurs when there is a sudden loss of blood flow to a part of the heart muscle. Heart attack always causes some permanent damage to the muscle, but the sooner treatment is given, the more muscle it is possible to save. If a heart attack damages a significant amount of the heart muscle, this can affect the pumping action of the heart (Mendis, 2005). In literatures, most of research focus on hospitalization and readmission of people having heart disease, which is corrective medicine (Agarwal, 2013). It is generally believed that predictive modeling of disease may reduce hospitalization of patients and it is a means of keeping patients away from returning to hospital unnecessary. The focus of this study is preventive medicine in that it will allow medical practitioners to ascertain the risk level of a patient developing myocardial infarction (heart attack).



(a)l



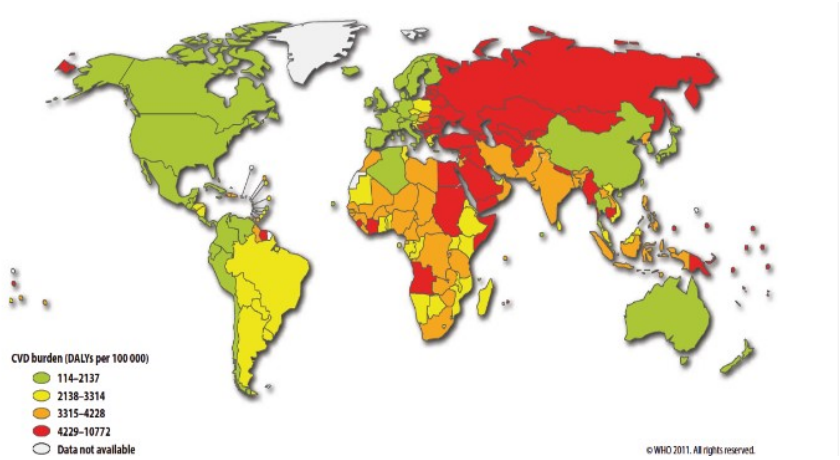
(b)

Figure 1 - World map showing the global distribution of CVD mortality rates in; (a) males, and (b) females (WHO, 2011)

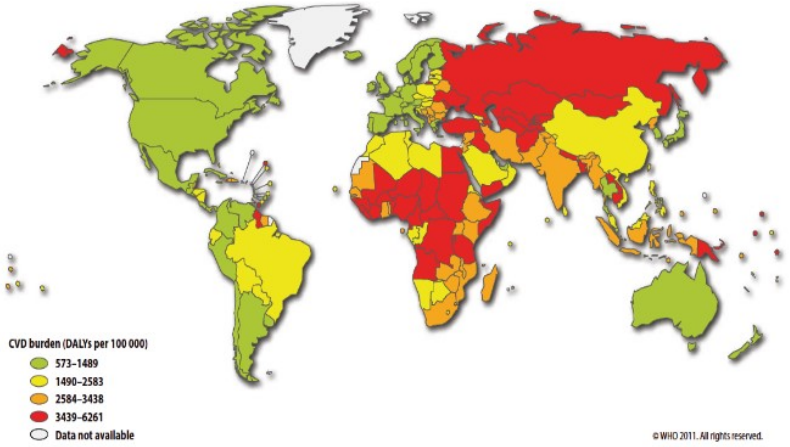
Nowadays, health care industry contains huge amount of heath care data, which contain hidden information. This hidden information is useful for making effective decisions (Dangare and Apte, 2012). Data mining has become a fundamental methodology for computing applications in medical informatics. It has great potential for exploring hidden patterns in a dataset of a medical domain, hence providing a user-oriented approach to novel and hidden patterns in a data (WHO, 2007; Abdullah and Rajalaxmi, 2012; and Thuraisingham (2000). Feature selection methods are unsupervised machine learning techniques used to identify relevant attribute in a dataset. It is important in identifying irrelevant and redundant attributes that exist within a dataset which may increase computational complexity and time (Yildirim, 2015; Hall, 1999). Feature selection methods are broadly classified as filter-based, wrapper-based and hybrid methods. Hybrid-based feature selection methods will be used to identify the most relevant variables, and these will be used in the predictive model for heart disease patients' classification using supervised machine learning techniques.

In many African countries, including Nigeria, the incidence and prevalence of heart disease is on the increase (Essien *et al*., 2014). The Word Health Organisation rates heart disease as one of the most important causes of premature death worldwide (WHO, 2012). This disease affects individuals in their peak and mid-life years, disrupting the future of the families that are dependent on them and undermining the development of the nations by depriving them of valuable human resources in their most productive years (WHO, 2002). In developing countries, heart disease tends to affect people at a younger age and thus could negatively affect the workforce and economic productivity (Leeder *et al*., 2009).

People in low- and middle-income countries often do not have the benefit of integrated primary healthcare programmes for early detection and treatment of people with risk factors as compared to high-income countries. As a result, this disease may be detected late; hence, many people die younger from the disease often in their most productive years. Although, there are various available treatments for prevention of sudden death in patients such as Implantable Cardioverter Defibrillators (ICD); however, these treatments are expensive and do not eliminate the risk of sudden death (Idowu *et al*., 2015).



(a)l



(b)

Figure 2 - World map showing the global distribution of the burden of CVDs (DALYs), in (a) males, and (b) females (WHO, 2011)

Cardiovascular disease events could be prevented or delayed if patients are aware of the disease risk early enough. Existing data mining models have been used to predict the existence (that is, presence or absence) of the disease rather than the risk of patients developing the disease. There is a need to determine the risk of a patient developing the disease to mitigate its incidence, hence this study. The aim of this study is to develop a model a model that will predict the risk of an individual developing heart attack. Its objectives are to formulate a predictive model to determine the likelihood of having heart attack in a patient, simulate the model formulated; and evaluate the performance of the model.

The rest of this paper is organized as follows. Section 2 discusses the relevant literature about the study while section 3 explains the materials and methods used for the study. Section 4 discusses the results obtained and section 4 concludes the paper.

2. RELATED WORKS

Chaurasia and Pal (2013) presented an approach to predict heart disease using data mining techniques. The dataset used was obtained from Urvine California Institute (UCI) and they applied three decision tree classifiers namely: Iterative Dichotomized 3 (ID3), Classification and regression tree (CART) and Decision Table (DT) for the diagnosis of patients with heart diseases or not. The analysis of the data and implementation of the model was carried out using Weka simulation environment tool. They used 10-fold cross validation to minimize any bias in the process, improve the efficiency of the process and evaluate the robustness of the classifier. Their classification result shows that CART classifier outperformed the other two with an accuracy of 83.49%; hence, it became their proposed method.

Taneja *et al.* (2014) developed a prediction model that can predict heart disease using data collected from Chandigarh in India consisting of 15 attributes. Filter based feature selection of information gain, gini index and chi squared were performed on the fifteen attributes (features) in order to obtain relevant features (eight features), thereby reducing the number of features the classification algorithm has to examine, and also reducing the errors from irrelevant features. 10-fold cross validation was adopted for randomly sampling the training and test data set. Decision tree, Neural Network and Naive Bayes were the classifiers used for the prediction model. The performance of the model was evaluated using the standard metrics of accuracy, precision, and recall. The J48 decision-tree outperformed the others.

Masethe and Masethe (2014) developed a model that can infer characteristics of predicted class from a combination of other data. The task of data mining in their

work is to build models for the prediction of the presence or absence of heart attack based on selected attributes for heart attack disease. They applied J48, Bayes Network and Naïve Bayes, Simple Cart and Reptree algorithm to classify and develop a model to diagnose heart attack in patient dataset. The dataset used in the study was collected from medical practitioners in South Africa. The dataset contains 108 records of patient with 11 attributes. These attributes are gender, age, chest pain, cholesterol, smoking, blood sugar, blood pressure, electrocardiographic (ECG), diet, alcohol, exercise level. For the analysis and implementation of the model WEKA was used. The stratified 10-fold cross validation was used to assess the performance of the algorithms in predicting a class. The work also determined which model gives the highest percentage of correct predictions for the diagnoses, of which J48, REPTREE and SIMPLE CART algorithm perform best.

Florence *et al*. (2014) predicted the occurrence of heart attacks in patient using neural network and decision tree. The dataset used was Acarth dataset which consist of six attributes such as age, sex, cardiac duration, cholesterol, signal level and possibility of attack, the final one being their class label. The output, which is the possibility of having heart attack, is the class label and it is either Yes or No. for evaluating the performance of their classifiers, 75% of the dataset was used for training and the remaining 25% for testing. The analysis of the data and implementation of the model for prediction was done using Rapid Miner Studio. They did not evaluate the robustness of the classifiers adopted in their model.

Nikhar and Karandikar (2016) worked on the prediction of the diagnosis of heart attack with a reduced number of attributes using Naive Bayes and Decision Tree. Records set with medical attributes were obtained from the Cleveland Heart Disease Database. With the help of the dataset, pattern significant to heart attack diagnosis were extracted - 19 attributes- these attributes are age, sex, chest pain, chest pain type, resting blood pressure on admission to the hospital (trestbps), chol, FBS, restecg, thalach (maximum heart rate achieved), exang (exercise induced angina) etc. The dataset consists of 303 records, when evaluating the performance of the classifier, 50% of the dataset was used for training and the remaining 50% was used for testing. Decision Tree outperformed Naïve Bayes classifier.

Shah *et al*. (2020) worked on the probability of developing heart disease in patients using machine learning techniques of Naïve Bayes, decision tree, K-nearest neighbor and random forest algorithm using WEKA tool. Existing dataset from the Cleveland database of UCI repository of heart disease patients was used. K-nearest neighbor gave the highest accuracy.

Tran-Duy et al. (2020) developed prediction models for classification of Cardiovascular Risk of Remote Indigenous Australians. They used clinical and demographic information on Indigenous people aged between 30 and 74 years without history of CVD events. They used Cox proportional hazard models to estimate 5-year CVD risk and the Harrell's c-statistic and the modified Hosmer-Lemeshow (mH-L) χ^2 statistic to assess the model discrimination and calibration, respectively. The study sample consisted of 1,583 individuals. The risk score consisted of sex, age, systolic blood pressure, diabetes mellitus, waist circumference, triglycerides, and albumin creatinine ratio. The bias-corrected c-statistic was 0.72 and the bias-corrected mH-L χ^2 was 12.01.

Limitations of Existing System

Many of the existing systems predict heart disease using attributes from complex test conducted in laboratories which may be expensive for low- or middle-income earners especially knowing fully well that the country in which the study is being conducted is categorized as developing country. Although Amin *et al.* (2013) proposed a hybrid model of genetic algorithm and ANN for predicting heart disease using risk factors not from complex test conducted in labs, as their input data. The datasets consist of 50 records (that is, risk factor data for 50 patients) and their result showed good classification accuracy. Their classification was binary based meaning that they aimed at predicting if a patient is healthy or heart diseased thereby predicting the occurrence of the existence of the disease rather than the occurrence of the severity of the disease.

3. MATERIAL AND METHODS

In this paper, a predictive model for heart attack risk level was proposed. The proposed predictive model is depicted in Figure 3. The model identifies the variables monitored in patients and those at the risk of developing the disease by the cardiologist at the study location. This was preceded by the collection of the dataset containing the identified variables for patients in the study location. The dataset collected from the hospital formed the basis of the historical dataset which contains various records of predictive parameters. Feature selection methods were used to identify the most relevant and important features among the features collected. The historical dataset containing the reduced feature set was divided into two – training dataset and testing dataset. The training and testing dataset were fed to the supervised machine learning algorithm proposed for this study using 10-fold cross validation evaluation method. The result of the performance of the combination of filter-based and each wrapper-based feature selection method together with the supervised machine learning algorithms were used to identify the most effective and efficient predictive model for heart attack risk prediction.

The data for 206 people were acquired from cardiologists at Obafemi Awolowo University Teaching Hospitals Complex, Ile-Ife, Osun State, Nigeria. Table 1 shows the identified important risk factors and their encoded values in brackets, which were used as input to the system. Data mining techniques were used in classifying the patient records as high risk, intermediate risk and low risk of developing the disease. The simulation of the project is carried out in Rapid Miner.

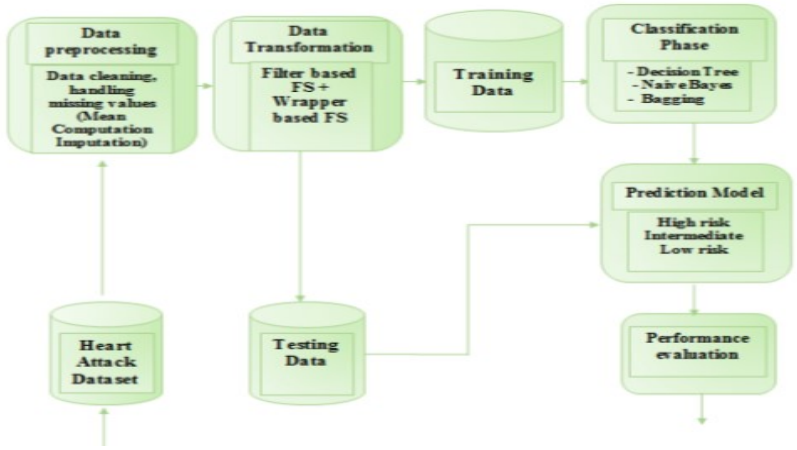


Figure 3 - Conceptual Framework of the Proposed Model

Table 1: Attributes Monitored and Units of Measure

Attribute	Type	Label
Age	Numeric	Years (yr)
Gender	Nominal	Female, Male
Family History	Nominal	Yes, No
Height	Numeric	Meters (m)
Weight	Numeric	Kilogram (kg)
BMI	Numeric	kgm^{-2}
Cholesterol Level	Numeric	$mmoL^{-1}$
Blood Pressure	Numeric	mmhg
Fasting Blood Sugar	Numeric	$mmol^{-1}$
Diabetes	Nominal	Yes, No
Exercise Level	Nominal	1, 2
Smoking	Nominal	Yes, No
Alcohol	Nominal	Yes, No
Associated Disease	Nominal	0, 1

The descriptions of the identified attributes used in this study are:

i. **Age** is the present age of the subject. It is measured in years and recorded using numeric figures.

ii. **Gender** refers to the sex of the subject. It is a variable with nominal value and without any unit of measure. It is recorded as "Male" or "Female".

iii. **History** specifies whether the parents (limited to first and second generations) once suffered from heart disease. It is a nominal variable, with no unit of measure. It is recorded as a "Yes" or "No"

iv. **Body Mass Index (BMI)** is obtained as the ratio of the body's weight (in Kilograms) to the square of the height (in meters) of the subject. It is measured in Kg/m2 and recorded as numeric value.

v. **Weight** is obtained as the weight of the patient's body which is measured in kilogram (kg)

vi. **Height is** obtained as how tall the patient is. It is measured in meters (m)

vii. **Systolic blood pressure** is the measure of the pressure (peak) in the arteries when the heart beats (when the heart muscles contract). It is measured in millimetre Mercury (mmHg) and recorded as numeric value.

viii. **Diastolic blood pressure** is the measure of the pressure (minimum) in the arteries between heartbeats (when the heart muscles relax between beats and refilling with blood). It is measured in millimetre Mercury (mmHg) and recorded as numeric value.

ix. **Cholesterol level** is measured in $mmol^{-1}$ and recorded as numeric value.

x. **Fasting Blood Sugar** is measured in m/mol and recorded as numeric value.

xi. **Diabetes** measure whether a subject has diabetes (which means "yes") or not (which means "no").

xii. **Exercise level** means whether the subject does enough exercise of at least 3 days for 30 minutes in a week or not. In this study, enough exercise of at least 3 days a week is denoted as "2" else denoted as "1"

xiii. **Smoking** means whether the subject smokes cigarette. It is recorded as nominal value.

xiv. **Alcohol intake** means whether a subject takes alcoholic drinks. It is recorded as nominal value.

xv. **Associated Disease** means whether the subject has diseases related to heart attack. The presence of other disease is denoted as "1" and its absence is denoted as "0".

For the purpose of handling the problem as a classification-based prediction model, the target classes (output variable) was determined using three labels, namely High, Low and Intermediate. These were assigned to each record to quantify each patient's measure of risk level. High refers to patients that have more than one risk factors plus the presence of associated disease. Intermediate refers to patients that have two or more risk factors only. Low refers to patients that have less than two risk factors. Algorithm 1 was used in assigning a target class to each patient's record using the World Health Organization (WHO) criteria of risk stratification.

Algorithm 1: Target Class Assignment

Start:
Let associated disease be X
Let number of risks be Y
If ((Y > 0) && (X = = Yes))
Then Target class = "High"
Else
if (0 ≤ Y ≤ 1))
Then Target class = "Low"
Else

Target Class = "Intermediate".
End if
End if
Stop

3.1. Data collection and Pre-Processing

Prior to the collection of the dataset, ethical approval was obtained from the Ethical and Research Committee of the OAUTHC, Ile-Ife, Nigeria. There was no need for consent forms since the patients were not required to partake in the study rather, data containing information about each patient excluding their personal information (e.g., names, address, hospital ID, contact number etc) were collected from health records unit in the Cardiac Care Clinic and General Out-Patient Department (GOPD). The data were stored electronically using Microsoft excel spread sheet. A total of two hundred and 6ix (206) records were collected data. The records in the data contained clinical information of ninety-one (91) males and one hundred and fifteen (115) females within the age range of 18–86. Figure 4 shows a screenshot of the data collected in an excel format.

For the purpose of handling the problem as a classification-based prediction model, the target classes (output variable) were determined using three labels, namely high, low, and intermediate were assigned to each record to quantify each patient's measure of risk level.

High: refers to patients that have more than one risk factors and the presence of associated disease.

Intermediate: refers to patients that have two or more risk factors only.

Low: refers to patients that have less than two risk factors.

Algorithm 1 was used in assigning a target class (High, Low, and Intermediate) to each patient's record using the WHO criteria of Risk Stratification.

After the collection of the data, some records in the data are found with missing values. Basically, the attributes found with missing values are BMI (1), cholesterol level (178), Blood pressure (4), Fasting blood sugar (161), Diabetes (4), exercise level (1), smoking (1), and Alcohol (1). Due to the negative effects large number of missing values could have on the prediction accuracy, the attributes with missing value above 50% of the entire dataset were removed. Thus, the cholesterol level and fasting blood sugar were deleted from the dataset, therefore reducing the attributes to thirteen. In this study, mean imputation method was used to populate the missing values. The decision of using mean imputation method in this study is based on the percentage of missing values in the dataset (<5%) and its overall effectiveness in improving the accuracy of classification algorithms. The algorithm for replacing the missing values is shown in Algorithm 2.

Algorithm 2 Mean Imputation Method

Let D = {X, Y} *// where D is the dataset with missing values*
//X = X_i ...X_n where i is the i^{th} attribute column of D with
//missing value(s)
//n is the number of //attributes
Let Value = V (X_{ij}) *//where V (X_{ij}) is the value of attribute i*
// in patient j

$$Mean\ of\ X_i \quad = m(X_i) \quad = \sum_{j=1}^{k} {}^{X_{ij}}/_{k} \qquad (1)$$

$$Mode\ of\ X_i = \quad \mathrm{M}(X_i) \quad = \max\left[\mathrm{V}(X_{ij})\right] \qquad (2)$$

Start:
For i = 1 to n *// for each attribute counting*
//from the first to the last attribute
For j = 1 to k *// for each records counting from the*
// first tuple to the last tuple
If (X_{ij} is null)
Do
If (X_i is nominal)
Return M(X_i);
Else
Return m(X_i)
End if

End if
Stop

AGE	GENDER	FAMILY H	HEIGHT	WEIGHT	BMI	CHOLESTE	SYSTOLIC	DIASTOLI	FBG	DIABETES	EXERCISE	SMOKING	ALCOHOL	ASSOCIATED
65	FEMALE	NO	1.72	70	23.66		140	70		NO	2	NO	NO	1
59	MALE	NO	1.69	69	24.16		140	80		NO	2	NO	YES	0
51	MALE	YES	1.43	61	29.83		120	90		NO	2	NO	NO	1
49	FEMALE	YES	1.61	72	27.78		120	80		NO	2	NO	NO	1
40	FEMALE	NO	1.59	52	20.57		110	70		NO	1	NO	NO	0
44	FEMALE	NO	1.71	68	23.3		140	90		NO	2	NO	NO	0
41	FEMALE	NO	1.65	60	22.04		110	70		NO	2	NO	NO	0
55	MALE	NO	1.77	78	24.9		130	90		YES	2	NO	NO	1
61	MALE	YES	1.63	72	27.1	3.6	130	70		NO	2	NO	YES	1
55	FEMALE	YES	1.54	73	30.78	1.3	120	60	4.8	NO	2	NO	NO	1
61	FEMALE	NO	1.77	72	22.98		120	70		YES	2	NO	NO	0
63	FEMALE	NO	1.73	71	23.72		110	70		NO	1	NO	NO	1
69	MALE	NO	1.74	65	21.47		120	60			2	NO	NO	1
68	MALE	NO	1.72	65	21.97		140	90		YES	2	NO	NO	1
66	FEMALE	NO	1.66	67	24.31		120	80		YES	2	NO	NO	1
45	MALE	NO	1.73	71	23.72		140	90		NO	2	NO	NO	0
48	MALE	NO	1.72	68	22.99		160	100		NO	2	NO	NO	0
67	FEMALE	NO	1.58	63	25.24		170	90	7.9	NO	1	NO	NO	1
45	MALE	NO	1.73	71	23.72		140	90		NO	2	NO	NO	0
47	FEMALE	YES	1.69	70	24.51		140	80		NO	2	NO	NO	0
73	FEMALE	NO	1.43	61	29.83		110	70		NO	1	NO	NO	1
68	FEMALE	YES	1.53	60	25.63	2.9	140	70	5.7	NO	1	NO	YES	1
45	FEMALE	YES	1.66	68	24.68		120	80		NO	2	NO	NO	0
41	MALE	YES	1.69	70	24.51		160	100		NO	2	NO	NO	0
43	MALE	NO	1.76	77	24.86		140	90		NO	2	NO	NO	0
60	FEMALE	YES	1.65	85	31.22		130	70	5.3	NO	1	NO	NO	1
43	FEMALE	NO	1.56	74	30.41	4	140	90		NO	1	NO	NO	1
41	FEMALE	NO	1.74	72	23.78		160	100		NO	2	NO	NO	1
68	MALE	NO	1.72	65	21.97		120	90		YES	2	NO	NO	1
68	FEMALE	YES	1.63	92	23.34		140	80	7.2	NO	1	NO	NO	1
45	FEMALE	NO	1.54	87	36.68	1.2	130	80		YES	1	NO	NO	1

Figure 4 - Sample Dataset in Microsoft Excel Worksheet

3.2. Data Transformation

Following the process of the identification and collection of the data needed for developing the predictive model, it was necessary to determine which set of variables are deemed more relevant effective for heart attack risk prediction. The basic algorithm for implementing the hybrid feature selection algorithm used in this study is by performing the filter-based feature selection using information gain ratio and running the result using the wrapper based feature selection, hence a hybrid feature selection. Hybrid technique performs just like wrapper approach; this method uses a filter method in the first pass to remove irrelative features and then a classifier specific wrapper method to further reduce the feature set. The attributes selected are shown in Table 2 and Table 3.

Table 2: Attributes Selected by Information Gain Ratio Feature Selection

Information Gain Ratio (IGR)
Associated Disease
Exercise level
BMI
Family History
Alcohol
Age
Systolic
Diabetes
Weight
Smoking

Table 3: Attributes Selected by Wrapper Feature Selection

Decision Tree based Wrapper Method (DTW)	Bagging based Wrapper Method (BW)	Naïve Bayes based Wrapper Method (NBW)
BMI	BMI	BMI
Systolic Blood Pressure	-	Systolic Blood Pressure
Diabetes	Diabetes	Diabetes
Exercise level	Exercise level	Exercise level
-	Smoking	Smoking
Alcohol	Alcohol	Alcohol
Associated Disease	Associated disease	Associated disease
Family History	Family History	Family History
Weight	Weight	-

4. RESULTS

To apply the data mining algorithms, the Rapid Miner version 7.5.001 was used to simulate the predictive model. For the purpose of developing the predictive model for determining the risk of a patient developing heart attack, stratified 10-fold cross validation method was employed. The performance measures considered to evaluate the algorithm are described and then, the obtained results are presented. The models were developed so that the different prediction model could be evaluated. These include the evaluation of data mining model of the whole features of the dataset, and the reduced dataset using feature selection methods of information gain ratio, decision tree wrapper-based feature selection, bagging wrapper-based feature selection and Naïve Bayes wrapper-based feature selection.

In order to determine these, four parameters of a confusion matrix must be identified from the results of predictions made by the classifier during model testing. These are: true positive (TP), true negative (TN), false positive (FP) and false negative (FP).

TP is the correct prediction of positive cases,

TN is the correct prediction of negative cases,

FP is the incorrect prediction of positive cases (negative predicted as positives).

FN is the incorrect prediction of negative cases.

Table 4 shows the description of the confusion matrix (CM), TPs are the values in the predicted class of an actual class; the total number of FNs for a class is the sum of values in the corresponding column (excluding the TP); The total number of FPs for a class is the sum of values in the corresponding row (excluding the TP); The total number of TNs for a certain class will be the sum of all columns and rows excluding that class column and row. The actual case of a confusion matrix is the sum of TP and FN (i.e. actual case = TP + FN). Taking for example:

Actual case for High (AC_H), AC_H = TP + FN; where FN is the sum of values in the corresponding column (i.e. FN = V_{HI} + V_{HL})

Actual case for Intermediate (AC_I) = TP + FN; where FN is the sum of values in the corresponding column (i.e. FN = V_{IH} + V_{IL})

Actual case for Low (AC_L), AC_L = TP + FN; where FN is the sum of values in the corresponding column (i.e. FN = V_{LH} + V_{LI})

Table 4: Description of a Confusion Matrix (CM)

Confusion Matrix		True		
		High	Intermediate	Low
Predicted	High	TP_H	V_{IH}	V_{LH}
	Intermediate	V_{HI}	TP_I	V_{LI}
	Low	V_{HL}	V_{IL}	TP_L

The Confusion matrix for the prediction models are shown in Table 5, Table 6, Table 7, Table 8, Table 9, Table 10, Table 11, Table 12, Table 13, Table 14, Table 15, Table 16, Table 17, Table 18 and Table 19.

Table 5: Confusion Matrix of Decision Tree Algorithm with 13 Attributes

	True H	True I	True L
Predicted H	15	41	1
Predicted I	111	11	2
Predicted L	0	5	20

Table 5 shows that 15 cases that are High Risk were correctly predicted as High Risk, 11 cases that were Intermediate were predicted as Intermediate and 20 cases that at Low Risk were correctly predicted as Low Risk. This shows that 46 cases were correctly predicted.

Table 6 shows the confusion matrix of Decision Tree algorithm with features selected from Information Gain Ratio. The table shows that there were 113 cases that are of High Risk predicted correctly as High Risk, 39 correctly predicted as Intermediate and 20 correctly predicted as Low Risk. This shows that there were 172 cases correctly predicted and 34 incorrectly classified.

Table 6: Confusion Matrix of Decision Tree Algorithm with Features Selected from Information Gain (IG) Ratio

	True I	True H	True L
Predicted H	113	14	1
Predicted I	11	39	2
Predicted L	2	4	20

Table 7 shows the result of classification of Decision Tree with features selected from Information Gain Ratio and Decision tree-based wrapper method. The table shows that 112 cases were correctly predicted as High Risk, 42 as Intermediate and 18 as Low Risk. This implies that 172 cases were correctly classified.

Table 7: Confusion Matrix of Decision Tree Algorithm with Features Selected from Information Gain Ratio (IG) and DT based Wrapper (DTW) Method

	True H	True I	True L
Predicted H	112	10	1
Predicted I	13	42	4
Predicted L	1	5	18

Table 8 gives the result of Decision Tree Algorithm with Features Selected from Information Gain Ratio (IG) and Bagging based Wrapper (BW) Method. It shows that 109 cases were correctly classified as High Risk, 37 as Intermediate and 17 as Low Risk, with a total of 163 cases correctly classified.

Table 8: Confusion Matrix of Decision Tree Algorithm with Features Selected from Information Gain Ratio (IG) and Bagging based Wrapper (BW) Method

	True H	True I	True L
Predicted H	109	15	2
Predicted I	16	37	4
Predicted L	1	5	17

Table 9 gives the Confusion Matrix of Decision Tree Algorithm with Features Selected from Information Gain Ratio (IG) and Naïve Bayes based Wrapper (NBW) Method. It shows that 106 cases were correctly classified as High Risk, 37 as Intermediate and 5 as Low Risk, with a total of 148 cases correctly classified and 48 incorrectly classified.

Table 9: Confusion Matrix of Decision Tree Algorithm with Features Selected from Information Gain Ratio (IG) and Naïve Bayes based Wrapper (NBW) Method

	True H	True I	True L
Predicted H	106	14	1
Predicted I	19	37	7
Predicted L	1	6	5

Table 10 gives Confusion Matrix of Bagging Algorithm with all the 13 attributes. It shows that 113 cases were correctly classified as High Risk, 43 as Intermediate Risk and 20 as Low Risk, with a total of 186 cases correctly classified and 29 incorrectly classified.

Table 10: Confusion Matrix of Bagging Algorithm with 13 Attributes

	True H	True I	True L
Predicted H	113	11	1
Predicted I	13	43	2
Predicted L	0	3	20

Table 11 gives the Confusion Matrix of Bagging Algorithm with Features Selected from Information Gain Ratio. It shows that 110 cases were correctly predicted as High Risk, 40 as Intermediate and 16 as Low Risk, with a total of 166 cases correctly classified and 30 cases incorrectly classified. Table 12 gives Confusion Matrix of Bagging Algorithm with Features Selected from Information Gain Ratio and DT based Wrapper Method. The table shows that 112 cases were correctly predicted as High Risk, 41 cases as Intermediate and 19 as Low Risk, with a total of 172 cases correctly predicted and 33 cases as incorrectly predicted.

Table 11: Confusion Matrix of Bagging Algorithm with Features Selected from Information Gain Ratio

	True H	True I	True L
Predicted H	110	11	1
Predicted I	15	40	6
Predicted L	1	6	16

Table 12: Confusion Matrix of Bagging Algorithm with Features Selected from Information Gain Ratio and DT based Wrapper Method

	True H	True I	True L
Predicted H	112	11	1
Predicted I	13	41	3
Predicted L	0	5	19

Table 13 gives the Confusion Matrix of Bagging Algorithm with Features Selected from Information Gain Ratio and Bagging based Wrapper Method. The result shows that 112 cases were correctly predicted as High Risk, 39 as Intermediate and 20 as Low Risk with a total of 171 cases correctly predicted and 35 incorrectly predicted. Table 14 gives Confusion Matrix of Bagging Algorithm with Features Selected from Information Gain Ratio and Naïve Bayes based Wrapper Method. It shows that 113cases were correctly predicted as High Risk, 39 cases as Intermediate Risk and 20 cases as Low Risk.

Table 13: Confusion Matrix of Bagging Algorithm with Features Selected from Information Gain Ratio and Bagging based Wrapper Method

	True H	True I	True L
Predicted H	112	12	1
Predicted I	14	39	2
Predicted L	0	6	20

Table 14: Confusion Matrix of Bagging Algorithm with Features Selected from Information Gain Ratio and Naïve Bayes based Wrapper Method

	True H	True I	True L
Predicted H	113	14	1
Predicted I	11	39	2
Predicted L	2	4	20

Table 15 gives Confusion Matrix of Naïve Bayes Algorithm with all the 13 Attributes. The table shows that 114 cases were correctly predicted as High Risk, 41 cases as Intermediate Risk and 15 cases as Low Risk with a total of 170 cases correctly predicted and 36 incorrectly predicted. Table 16 gives the Confusion Matrix of Naive Bayes Algorithm with Features Selected from Information Gain Ratio. It shows that 114 cases were correctly predicted as High Risk, 41 as Intermediate Risk and 15 cases as Low Risk with a total of 170 cases correctly predicted and 36 incorrectly predicted.

Table 16 gives the Confusion Matrix of Naive Bayes Algorithm with Features Selected from Information Gain Ratio. The table shows that 114 cases were correctly predicted as High Risk, 41 cases as Intermediate and 16 as Low Risk with a total of 171 cases correctly predicted and 35 incorrectly predicted. Table 17 gives the Confusion

Matrix of Naïve Bayes Algorithm with Features Selected from Information Gain Ratio and DT based Wrapper Method. It shows that 114 cases were correctly predicted as High Risk, 38 cases as Intermediate and 15 cases as Low Risk with a total of 167 cases correctly predicted.

Table 15: Confusion Matrix of Naïve Bayes Algorithm with 13 Attributes

	True H	True I	True L
Predicted H	114	13	1
Predicted I	12	41	7
Predicted L	0	3	15

Table 16: Confusion Matrix of Naive Bayes Algorithm with Features Selected from Information Gain Ratio

	True H	True I	True L
Predicted H	114	13	1
Predicted I	12	41	6
Predicted L	0	3	16

Table 17: Confusion Matrix of Naïve Bayes Algorithm with Features Selected from Information Gain Ratio and DT based Wrapper Method

	True H	True I	True L
Predicted H	114	14	1
Predicted I	12	38	7
Predicted L	0	5	15

Table 18 gives the Confusion Matrix of Naïve Bayes Algorithm with Features Selected from Information Gain Ratio and Bagging based Wrapper Method. It shows that 115 cases were correctly predicted as High Risk, 42 as Intermediate and 15 as Low Risk with a total of 172 cases correctly predicted and 34 cases incorrectly predicted. Table 19 gives Confusion Matrix of Naïve Bayes Algorithm with Features Selected from Information Gain Ratio and Naïve Bayes based Wrapper Method. The table shows that 114 cases were correctly predicted as High Risk, 48 cases as Intermediate Risk and 19 cases as Low Risk with a total of 181 cases correctly predicted and25 cases incorrectly predicted.

Table 18: Confusion Matrix of Naïve Bayes Algorithm with Features Selected from Information Gain Ratio and Bagging based Wrapper Method

	True H	True I	True L
Predicted H	115	13	1
Predicted I	11	42	7
Predicted L	0	2	15

Table 19: Confusion Matrix of Naïve Bayes Algorithm with Features Selected from Information Gain Ratio and Naïve Bayes based Wrapper Method

	True H	True I	True L
Predicted H	114	6	3
Predicted I	12	48	1
Predicted L	0	3	19

4.1. Performance Evaluation

In this study, the effect of feature selection on classification performance was evaluated. For evaluating the predictive models, accuracy and sensitivity were used as performance metrics.

Sensitivity, also known as *true positive rate or recall* is the proportion of actual positive cases that were correctly predicted positive by the model.

$$Sensitivity = \frac{TP}{TP+FN} \quad (3)$$

Accuracy: is the ratio of correctly classified samples to the total number of tested samples.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (4)$$

The performance measures for the prediction model executed on the whole set of features and reduced set of features are represented in Table 20. The table shows the correct classification, accuracy, mean recall and sensitivity values using Decision Tree, Bagging and Naïve Bayes classifiers when (i) all the features were selected, denoted as ALL, (ii) with only features selected from information gain ratio (IGR), (iii) with features selected from Information Gain Ratio and Decision Tree-based Wrapper Method (IGR+DTW), (iv) with features selected from Information Gain Ratio and Bagging-based Wrapper Method (IGR+BW) and (v) with features selected from Information Gain Ratio and Naïve Bayes-based Wrapper Method (IGR+NBW). Bagging achieved a better accuracy, which is above 83.50% when executed on the whole attributes. Naïve Bayes offers competitive accuracy of 82.52% but decision tree accuracy is considerably lower. The results obtained showed that training and testing the Naïve Bayes algorithm with features selected from information *gain* and *Naïve Bayes wrapper-based Feature selection* has the highest accuracy and recall of 87.86% and 85.77% respectively, hence, the best proposed prediction model for predicting heart attack in patient. Correct classification values for Naïve Bayes are shown in Figure 2 while the accuracy values are as depicted in Figure 3 and Mean Recall values are as depicted in Figure 4.

Table 20: Description of the Result of Each Model Using 10 Fold Cross-validation

Dataset		ALL	IGR	IG R+DTW	IGR+ BW	IGR +NBW
Decision Tree	Correct classification	163	158	176	172	172
	Accuracy	79.13%	76.70%	85.44%	83.50%	83.50%
	Mean Recall	75.11%	71.42%	84.03%	82.33%	81.69%
Bagging	Correct classification	172	166	173	171	172
	Accuracy	83.50%	80.58%	83.98%	83.01%	83.50%
	Mean Recall	80.28%	75.68%	81.41%	81.42%	81.69%
Naïve Bayes	Correct classification	170	171	167	172	181
	Accuracy	82.52%	83.01%	81.07%	83.50%	87.86%
	Mean Recall	75.875	77.32%	74.12%	76.72%	85.77%

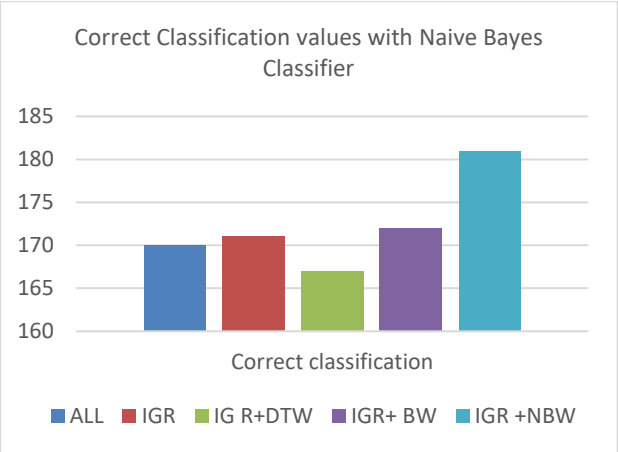


Figure 2 Correct Classification values with Naïve Bayes

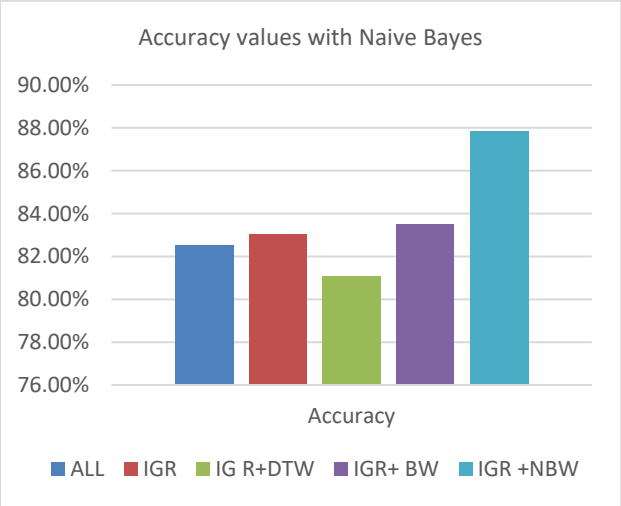


Figure 3 Accuracy values with Naïve Bayes

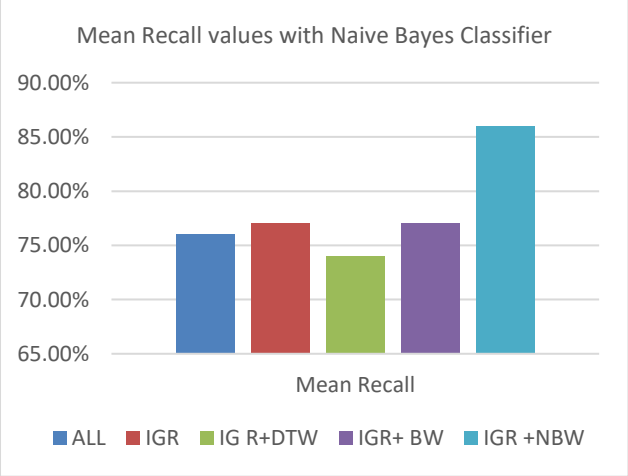


Figure 4 Mean Recall values with Naïve Bayes

5. CONCLUSION

In conclusion, this study has proposed a predictive model which combines filter-based feature selection and wrapper-based feature selection together with supervised machine learning algorithm. The predictive model has shown the important attributes relevant for the prediction of heart attack risk thus reducing the computational complexity of the model with an increased performance. Therefore, this study could be integrated into Health Information System (HIS) which captures, stores, and manages clinical and demographic information about patients. Such information can be fed to the heart attack risk prediction model, thereby providing aid to physicians, and improving clinical decisions affecting heart attack. The resulting model can also be used by junior cardiologist to help identify patients who needed special attention by screening out patients who have a high level of developing the disease and transferring those patients to senior cardiologist for further analysis.

REFERENCES

Abdullah A. S. and Rajalaxmi R. R. (2012). A Data Mining Model for Predicting the Coronary Heart Disease using Random Forest Classifier. International Conference on Recent Trends in Computational Methods, Communication and Control, Vol. 3, Issue 1, pp. 22-25.

Agarwal V., Briasoulis A. and Messerli F. H. (2013). Effects of renin-angiotensin system blockade on mortality and hospitalization in heart failure with preserved ejection fraction. Heart Failure Reviews, Vol. 4, Issue 3, pp. 429-437.

Amin S. U., Agarwal K. and Beg R. (2013). Genetic Neural Network Based Data Mining in Prediction of Heart Disease Using Risk Factors. Proceedings of 2013 IEEE Conference on Information and Communication Technologies (ICT 2013), Vol. 1, Issue 1 pp. 1228-1231.

Chaurasia V. and Pal S. (2013). Early Prediction of Heart Diseases Using Data Mining Techniques," Caribbean Journal of Science and Technology, Vol. 1, pp. 208-217.

Dangare C. S. and Apte S. S. (2012). A Data Mining Approach for Prediction of Heart Disease using Neural Networks. International Journal of Computer Engineering & Technology, Vol. 3, Issue 3, pp. 20-30.

Florence S., Bhuvaneswari N. G., Amma C., Annapoorani G. and Malathi K. (2014). Predicting the Risk of Heart Attacks using Neural Network and Decision Tree. International Journal of Innovative Research in Computer and Communication Engineering, Vol. 2, Issue11, pp. 7025-7030.

Hall, M. A. (1999). Correlation-based feature selection for machine learning. Unpublished Ph.D thesis, Department of Computer Science, University of Waikato, Hamilton, New Zealand.

Joshi S. and Nair M. K. (2015). Prediction of Heart Disease Using Classification Based Data Mining Techniques. Computational Intelligence in Data Mining, Vol. 2 Issue 1, pp. 503-511.

Mackay J. and Mensah G. A. (2004). The Atlas of Heart Disease and Stroke. Cardiovascular Disease, World Health Organization. [Online] Available at: https://www.who.int/cardiovascular_diseases/resources/atlas/en/ Accessed: 25th July, 2020.

Masethe H. D. and Masethe M. A. (2014). Prediction of Heart Disease using Classification Algorithms. Proceedings of the World Congress on Engineering and Computer Science, Vol. 2, Issue 1, pp. 25-29.

Mendis S., Abegunde D., Yusuf S., Ebrahim S., Shaper G., Ghannem H. and Shengelia B. (2005). "WHO study on Prevention of Recurrences of Myocardial Infarction and Stroke (WHO-PREMISE)," Bulletin of the World Health Organization, Vol. 83, Issue 3, pp. 820-829.

Nikhar S. and Karandikar A. M. (2016). Prediction of Heart Disease Using Machine Learning Algorithms. International

Journal of Advanced Engineering, Management and Science (IJAEMS), Vol. 2, Issue 6, pp. 617-621.
Oguanobi N. I., Ejim, E. C., Onwubere, B. J., Ike, S. O., Anisiuba, B. C., Ikeh, V. O. and Aneke, E. O. (2013). Pattern of cardiovascular disease amongst medical admissions in a regional teaching hospital in Southeastern Nigeria. Nigerian Journal of Cardiology, 10(2): 77.
Roger V. L., Go A. S., Lloyd-Jones D. M., Benjamin E. J., Berry J. D., Borden W. B. and Fullerton H. J. (2012). Heart disease and stroke statistics - 2012 update. A report from the American Heart Association, Circulation, Vol. 125, Issue 1, pp. 2-220.
Shah D., Patel S. and Bharti S.K. (2020). Heart Disease Prediction using Machine Learning Techniques. Springer Nature Computer Science Journal, 1:345 pp 1-6.
Soni J., Ansari U. and Sharma D. (2011). Intelligent and Effective Heart Disease Prediction System using Weighted Associative Classifiers. International Journal on Computer Science and Engineering, Vol. 3 Issue 6, pp. 43 – 48.
Taneja T., Gaurav G. and Sethi A. (2014). Study of classifiers in Data Mining. International Journal of Comput*er Science and Mobile Computing*, Vol. *3, Issue* 9, pp. 263-269.
Thuraisingham B. (2000). A primer for understanding and applying data mining. Institute of Electrical Electronics Engineers, IT Professional, Vol. 2, Issue 1, pp. 28-31.
Tran-Duy A., McDermott R., Knight J., Hua X., Elizabeth L. M., Arabena K., Palmer A. and Clarke P. M. (2020). Development and Use of Prediction Models for Classification of cardiovascular Risk of Remote Indigenous Australians. Heart, Lung and Circulation 29, 374-383.
WHO (2007). A Safer Future: Global Public Health Security in the 21st Century. The World Health Report. [Online] Available at: https://www.who.int/whr/2007/whr07_en.pdf. Accessed: 25th July, 2020.
WHO (2009). Commission on Macroeconomics and Health. Improving health outcomes of the poor. Report of working group 5 Geneva, Switzerland. [Online] Available at: http://who.int/iris/bitstream/10665/42488/1/9241590130.pdf. Accessed: 25th July, 2020.
Yildirim P. (2015). Filter-based feature selection methods for prediction of risks in hepatitis disease. International Journal of Machine Learning and Computing, Vol. 5 Issue 4, pp. 258.