

Predictive Model for the Classification of Hypertension Risk Using Decision Trees Algorithm

Idowu Peter Adebayo

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

Email address:

paidowu@oauife.edu.ng

To cite this article:

Idowu Peter Adebayo. Predictive Model for the Classification of Hypertension Risk Using Decision Trees Algorithm. *American Journal of Mathematical and Computer Modelling*. Vol. 2, No. 2, 2017, pp. 48-59. doi: 10.11648/j.ajmcm.20170202.12

Received: December 8, 2016; **Accepted:** December 19, 2016; **Published:** February 24, 2017

Abstract: This study is focused with the development of a predictive model for the classification of the risk of hypertension among Nigerians using decision trees algorithms based on historical information elicited about the risk of hypertension among selected respondents in southwestern Nigeria. Following the identification of the risk factors of hypertension from experienced cardiologists, structured questionnaires were used to collect information about the risk factors and the associated risk of hypertension from selected respondents. The predictive model was formulated using two (2) decision trees algorithms, namely: C4.5 and ID3 based on the information collected. The predictive model was simulated using the Waikato Environment for Knowledge Analysis (WEKA) using the 10-fold cross validation technique for model training and testing. The results revealed that the decision trees algorithms selected some risk factors among those identified as most predictive for the risk of hypertension based on the information inferred from the dataset collected. The variables were used by the decision trees algorithm to deduce the decision trees that were used to infer the risk of hypertension based on the values of the identified risk factors. The ID3 with an accuracy of 100% outperformed the C4.5 which showed an accuracy of 86.36%. The variables identified by the algorithms can help assist cardiologists concentrate on a smaller yet important set of risk factors for identifying the risk of hypertension using rules derived from the path along the decision trees based on the value of the risk factors of the individual.

Keywords: Hypertension Risk Factors, ID3, C4.5, Prediction, Classification, Decision Trees

1. Introduction

Data mining is a process of discovering meaningful useful information in large data repositories. Data mining can discover valuable but hidden knowledge from databases especially those used in storing health-related information about diseases affecting patients [1]. Clinical decisions are often made based on doctors' intuition and experience rather than on the knowledge rich data hidden in the database. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients [2]. The integration of clinical decision support with computer based patient records could reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome [3]. This suggestion is promising as data modeling and analysis tools like data mining have the potential to generate a knowledge-rich environment which can help to significantly improve the

quality of clinical decisions. Predictive research aims at predicting future events or an outcome based on patterns within a set of variables and has become increasingly popular in medical research [4] [5]. Accurate predictive models can inform patients and physicians about the future course of an illness or the risk of developing illness and thereby help guide decisions on screening and/or treatment [6].

Hypertension, or high blood pressure, is dangerous because it can lead to strokes, heart attacks, heart failure, or kidney disease and many more disease ailments [7]. It occurs when the person's mean arterial pressure is greater than the upper range of accepted normal pressure. According to [8], a mean arterial pressure of 110mmHg is considered hypertensive. This level of mean pressure occurs when the diastolic blood pressure is greater than 80mmHg (120/80mmHg) and the systolic pressure is greater than 120mmHg. Many cardiologists consider systolic pressure of 160mmHg and diastolic pressure of 100mmHg (160/100mmHg) to be hypertensive [9]. Several community

surveys indicate that the prevalence of hypertension in Nigeria has increased from 11.2% in the 1990s to 27.9% in 2010 within the Niger Delta [10] and 22.6% in 2009 among a suburban Christian community in south-west Nigeria. NCDs are also currently responsible for at least 20% of all deaths in Nigeria [11] and constitute up to 60% of the patients admitted into the medical wards of most tertiary hospitals in Nigeria [12]. The poor treatment outcome for non-communicable diseases is however recognized globally, and has prompted the WHO to propose a paradigm shift in health care delivery, in favor of preventive and more proactive healthcare [3] [13]. About 45 years ago, Omran had in a series of articles proposing the epidemiological transition theory [14]. In this theory, he predicted the displacement of infectious diseases by NCD, as major causes of morbidity and mortality, as a community or country develops. This theory has since been confirmed in most countries of the world, including Nigeria [13], [10], [12].

Uncontrolled hypertension is associated with serious end-organ damage such as heart disease, stroke, renal disease and blindness [15], [16], [17]. These serious complications can be prevented by adequate blood pressure control [18], [19]. The prevention and control of hypertension has not received due attention in many developing countries although it is one of the most modifiable risk factors for cardiovascular disease. Reliable epidemiologic data are useful for the design and implementation of effective strategies for the prevention and control of hypertension. The associated risk factors of hypertension include genetic or strong family history and other factors which include increasing age, obesity, smoking/use of tobacco, diabetes mellitus, dietary consumption of high salt content and saturated fat, sedentary lifestyle, stressful life, poor sleep and pregnancy [20], [21]. Among an eastern Nigeria population, alcohol consumption rate among adults was observed to be as high as 55.8% [22]. Symptoms of hypertension observed include headaches (20.3%), palpitations (12.4%), chest pain (7.1%), blurred vision (7.1%), breathlessness, dizzy spells, tinnitus, vertigo and insomnia [23]. Also, many patients only become aware of their hypertensive status after the development of target organ damage like stroke, hypertensive retinopathy, ischemic heart disease, congestive heart failure, peripheral vascular disease or chronic kidney disease [24]. Management of hypertension involves conducting necessary investigations and treatment. As Nigeria strives to become one of the twenty leading economies by 2020, there is urgent need to develop and implement control strategies to curb this health menace that is claiming the lives of millions Nigerians on yearly basis. As reported by two prominent National dailies in the country, more than 20 million cases of hypertension were estimated in 2010, affecting one in three men and one in four women and the figure is expected to rise to 39 million cases by 2030 [25].

Despite the differences and clashes in approaches, the health sector has more need for data mining today. There are several arguments that could be advanced to support the use of data mining in the health sector, covering not just concerns

of public health but also the private health sector. There is a wealth of knowledge to be gained from computerized health records. Yet the overwhelming bulk of data stored in these databases makes it extremely difficult, if not impossible, for humans to sift through it and discover knowledge [26]. In fact, some experts believe that medical breakthroughs have slowed down, attributing this to the prohibitive scale and complexity of present-day medical information. Computers and data mining are best-suited for this purpose [27]. When medical institutions apply data mining on their existing data, they can discover new, useful and potentially life-saving knowledge that otherwise would have remained inert in their databases, such safety issues could be flagged and addressed by hospital management and government regulators. Cao *et al* [28] described the use of data mining as a tool to aid in monitoring trends in the clinical trials of cancer vaccines. By using data mining and visualization, medical experts could find patterns and anomalies better than just looking at a set of tabulated data. Health experts have also begun to look at how to apply data mining for early detection and management of pandemics.

Hypertension cases are continuously on the increase in Nigeria and the number of hospital admissions and related mortality and morbidity of hypertension is also on a steady increase. Hypertension is constantly becoming a serious issue and sending many young Nigerian men and women to their grave. Unfortunately, majority of Nigerians do not realize the presence of hypertension unless an associated cardiovascular disease in most case is observed. Presently, there is not preventive mechanism that is aimed at avoiding or at least reducing the onset of hypertension or its related diseases. There is a need for a model that will capture details relevant to the assessment of the risk of hypertension so as to reduce the untimely deaths associated with hypertension prevalence in Nigeria; which is the focus of this paper to use decision trees algorithm to identify variables important for predicting the risk of hypertension and use them to develop a predictive model for the risk of hypertension in non-hypertensive patients. Nigeria has one of the highest burden of prevalence of hypertension in Sub-Sahara Africa and the largest percentage of these are due to the exposure to the risk factors of high blood pressure such as age, obesity, sex and lifestyle habits like high salt in-take, smoking, lack of physical activity, alcohol and stress. Data mining techniques can help identify relationships and activities that were not observed by medical health experts needed in improving the decision affecting the early detection of hypertension thereby reducing the associated deaths. Early detection of the risk of hypertension will help provide potential individuals of alternative pattern to lifestyle and dietary needs so as to avert the onset of hypertension and/or its related cardiovascular diseases.

2. Related Works

A number of related works have been done in the area of the application of data mining techniques to elicit knowledge

from health-related data regarding the risk of diseases especially in the area of cardiovascular diseases (CVD). A number of such papers are summarized in the following paragraphs.

Mayilvaganan and Rajeswari [29] proposed a human blood pressure classification analysis model using fuzzy logic control system. The model used a number of factors that were identified as required in monitoring the blood pressure for the purpose of early detection of slight increase in blood pressure. The model was implemented using fuzzy membership functions but the model was not validated using live datasets. This is a limitation to the work as a live dataset used in validating the model would have provided a better understanding of the model's performance upon deployment. The model's strength lie in the use of rules elicited from expert cardiologist which although may be subjective based on individual experiences.

Waghmare [30] worked on the prediction of the occurrence of non-communicable diseases (NCD) depending on lifestyle habits using data mining techniques. The datasets of 303 employees working in a public company in Mumbai, India from 2009 to 2011 was used to develop the model. The k-means clustering algorithm was used to develop the model for predicting the risk of developing NCDs using information about lifestyle habits. Diabetes, hypertension and dyslipidaemia were the NCDs considered in the study while the risk factors considered were limited to lifestyle habits whereas other risk factors exist which are liable to improve the performance of the prediction model

Ephzibah [8] proposed a fuzzy and genetic hybrid algorithm model for the prediction of diabetes. This work was applied to minimize cost and maximize accuracy. With the help of genetic algorithm, the computation cost decreases and the classification performance increases. It shows that by applying this principle within a fuzzy logic framework can significantly improve the mechanism's performance to

diagnose patients having diabetes. The classification accuracies of these datasets were obtained by k-fold Cross-validation. It achieves accuracy values of 84.24%.

Duen-Yian *et al.* [16] developed a predictive model for cerebrovascular disease using data mining using 493 valid samples and contains 29 attributes which includes blood test, physical exam results and diagnosis results. It adopted three classification algorithms, decision tree, and Bayesian classifier and back propagation neural network, to construct classification models. After applying the classification rules decision tree showed 99.59% accuracy and constructed a classification model with stable classification efficiency. The result indicated that the decision tree is the best classification algorithm when compared with other algorithms.

3. Materials and Methods

Figure 1 shows a description of the process involved in the methodology design of this study used in the development of the predictive model for hypertension. Following the review of related works of literature in the body of knowledge of hypertension and the factors related to its risk, a number of variables (risk factors) were identified. The identified risk factors of hypertension were validated by an expert cardiologist with more than 20 years' experience in medical practice before the instrument of data collection was constructed alongside the identification of respondents. The selected data collection instrument for this study is the questionnaire due to the problem associated with the unavailability of data related to risk of hypertension but for those with the disease. The questionnaire was constructed with the purpose of acquiring information about the associated risk of hypertension identified by the expert cardiologist from individuals with no hypertension history and from those with disease associated with the risk of hypertension like: diabetes, renal diseases etc.

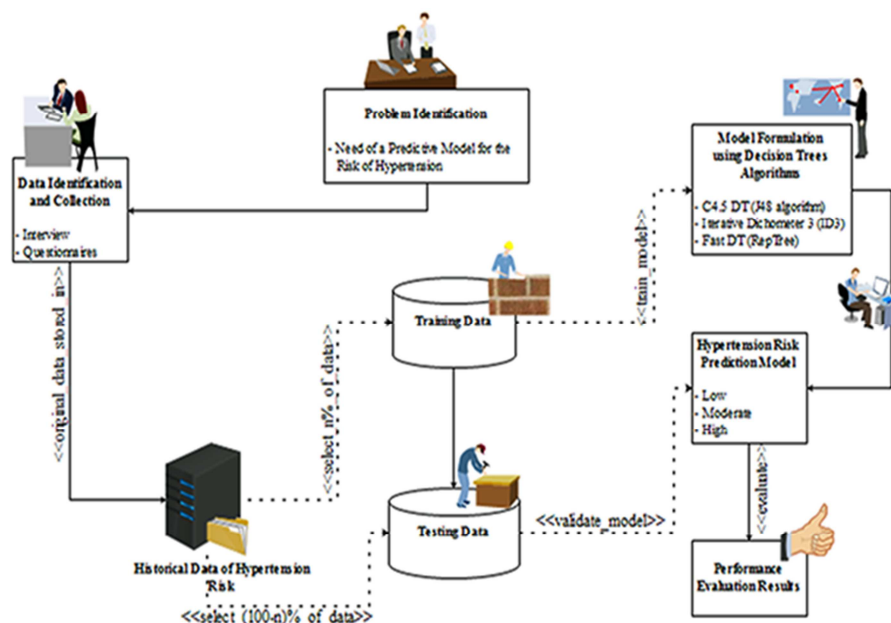


Figure 1. Research Design of the Study.

a Data identification and collection

The constructed questionnaire consisted of three (3) sections, namely sections A, B and C. Section A of the questionnaire consisted of information relevant to the individual's demographic information, namely: sex, age, education, occupation, marital status, job position, area of residence and ethnicity. Section B of the questionnaire consisted of information relating to the risk factors of the risk hypertension from the individual respondent. The information required includes: family history, body mass index (BMI), frequency of exercise, marital/domestic issues, stress in office, hours sent at work, smoking and length, alcohol consumption, brand and duration and salt intake. Section C consists of the doctor's comments; this space is left free for the doctor to provide his comment on the associated risk of hypertension based on the information provided on each questionnaire. It is important to state that the comments provided by the cardiologist is subjective to his own experience in medical practice and may not be a true representation of the generic risk of hypertension in Nigeria.

b Formulation of Predictive Model

Following the identification and validation of variables relevant to the risk of hypertension and the collection of historical explaining the relationship between the identified risk factors and their respective risk for each record of individuals, the predictive model for the risk of hypertension was formulated using the decision trees algorithm. In this study, supervised machine learning algorithms was used in formulating the predictive model since the pattern explaining the relationship between the identified factors (input variables) and the respective risk of hypertension (the target variable) was required. The identified pattern can then be converted into a set of rules that can help assist cardiologist make informed decisions about the risk of hypertension in Nigerians. For any supervised machine learning algorithm proposed for the formulation of a predictive model, a mapping function can be used to easily express the general expression for the formulation of the predictive model for the risk of hypertension – this is as a result that most machine learning algorithms are black-box models which use evaluators and not power series/polynomial equations.

The historical dataset S which consists of the records of individuals containing fields representing the set of risk factors (i number of input variables for j individuals), X_{ij} alongside the respective target variable (risk of hypertension) represented by the variable Y_j – the risk of hypertension for the jth individual in the j records of data collected from the hospital selected for the study. Equation 1 shows the mapping function that describes the relationship between the risk factors and the target class – risk of hypertension.

$$\varphi: X \rightarrow Y; \quad (1)$$

$$\text{defined as: } \varphi(X) = Y$$

The mapping φ was used to represent the predictive model for hypertension risk maps the risk factors of each individual

to their respective risk of hypertension according to equation 2.

$$\varphi(X) = \begin{cases} \text{Low risk} \\ \text{Moderate risk} \\ \text{High risk} \end{cases} \quad (2)$$

c Decision trees (DT) algorithm used

The formulation of the predictive model for hypertension risk in individuals was proposed using decision trees algorithm for the classification of the risk of hypertension as either of low, moderate and high risk given the values/labels of the identified risk factors (nodes) used in the development of a hierarchical tree structure using a splitting criteria. Each interior node (starting from the root/parent node) of the decision tree represents the attributes (important risk factors of hypertension risk) with edges that correspond to the values/labels of each attributes leading to a child node (another attribute conditional to the value of the parent node) at the bottom; this process continues for each subsequent values of the attributes until the leaf is reached - the terminal node also representing the target class (risk of hypertension – low, moderate or high).

During the training process of model development using the historical dataset collected, the pattern is learned by the tree by splitting the training dataset into subsets based on an attribute value test for each input variables; the process is repeated on each derived subset in a recursive manner called recursive partitioning. The recursion is completed when the subset at a node has all the same value of the target class, or when splitting no longer adds value to the predictions. This is also called the Top-down induction of trees which is an example of the greedy algorithm also called divide-and-conquer.

The training data set, S is a set containing S_1, S_2, \dots, S_j of already classified samples of the records of individuals consisting of the values of their risk factors, $X = \{X_1, X_2, \dots, X_i\}$ alongside the risk of hypertension, $Y = \{\text{low}, \text{moderate}, \text{high}\}$ such that, $S = (X, Y)$ for all individuals from 1 to j. The decision trees algorithms used to develop the predictive model for the risk of hypertension in individuals alongside their respective criteria are presented as follows:

- ID3 (Iterative Dichotomiser) decision trees algorithm

ID3 builds the DT from a set of training dataset, S by making it the root node. For each iteration of the algorithm, ID3 iterates through every unused attribute, X (risk factor) of the set S and calculates the *entropy*, $H(S)$ (or *information gain*, $IG(X)$) of that attribute (risk factor). It then selected the attribute which has the smallest entropy (or largest IG) value. The set, S was then spitted by the selected attribute into labels (e.g. present age is below 20, between 21 and 30, between 31 and 40, above 40) to produce subsets of the data. The algorithm then continued to recourse on each subset, considering only attributes never used before. Throughout the ID3 algorithm, the DT is constructed with each non-terminal node representing the selected attribute (risk factor) on which the data was split and the terminal nodes represented the

class label (risk of hypertension) of the final subset of the branch. Equations (3) and (4) show the two (2) metrics used by the id3 DT algorithm in constructing the DT for hypertension risk in pregnant women.

Entropy $H(S)$ is a measure of the amount of uncertainty in the data S while the information gain $IG(X)$ is the measure of the difference in entropy from before and after the dataset S was split on an attribute X using a subset of labels T .

$$\text{Entropy}, H(X) = -\sum_{x \in X} p(x) \log_2 p(x) \quad (3)$$

$$\text{Information Gain } IG(X) = H(X) - \sum_{t \in T} p(t)H(t) \quad (4)$$

• C4.5 decision trees algorithm

C4.5 algorithm builds DT from a set of training dataset, S the same way as ID3, using the *information entropy*. At each node of the tree, C4.5 chose the attribute (risk factor) of the training data that most effectively splits its set of samples into subsets enriched in one class or the other (risk of hypertension). The splitting criteria used is then normalized *information gain* (difference in entropy). The attribute with the highest normalized information gain is chosen to make the decision. Then C4.5 recurs on the smaller sub-list of attributes. The C4.5 algorithm has a few base cases:

- a All the samples in the list belong to the same class. When this happens, it simply creates a leaf node for the decision tree saying to choose that class.
- b None of the features provide any information gain. In this case, C4.5 creates a decision node higher up the tree using the expected value of the class.
- c Instance of previously-unseen class encountered. Again, C4.5 creates a decision node higher up the tree using the expected value.

The two criteria used by the C4.5 decision trees in developing its DT are presented in equations (5) and (6) as the information gain and the split criteria respectively.

$$IG(X) = H(X) - \sum_{t \in T} \frac{|t|}{|X|} \cdot H(t) \quad (5)$$

Where:

$$H(X) = -\sum_{t \in T} \frac{|t, X|}{|X|} \cdot \log_2 \frac{|t, X|}{|X|}$$

$$\text{Split}(T) = -\sum_{t \in T} \frac{|t|}{|X|} \cdot \log_2 \frac{|t|}{|X|} \quad (6)$$

d Model simulation process and environment

The Waikato Environment for Knowledge Analysis (WEKA[®]) software – a suite of machine learning algorithms was used as the simulation environment for the development of the predictive model. The software is freely available at <http://www.cs.waikato.ac.nz/ml/weka>. The system was written using object-oriented language, Java. There are several different levels at which Weka can be used. Weka provides implantations of state-of-the-art data mining and machine learning algorithms. Weka contains modules for data preprocessing, classification, clustering and association rule extraction for market basket analysis.

The main features of Weka include:

- a 49 data preprocessing tools;
- b 76 classification/regression algorithms;
- c 8 clustering algorithms;
- d 15 attribute/subset evaluators + 10 search algorithms for feature selection;
- e 3 algorithms for finding association rules; and
- f 3 graphical user interfaces, namely:
 - g The Explorer for exploratory data analysis;
 - h The Experimenter for experimental environment; and
 - i The Knowledge Flow, a new process model inspired interface.

The datasets were subjected to 10-fold cross validation using the two (2) selected decision trees learning algorithms, namely: C4.5 implemented as the J48 algorithm on WEKA and the ID3 algorithm. Before subjecting the historical datasets containing the values of the risk factors alongside the risk of hypertension for each respondent's record in the original dataset; there was the need of storing the dataset according to the default format for data representation needed for data mining tasks on the Weka environment. The default file type is called the attribute relation file format (.arff). the arff file type stores three category of data: the first defining the title of the relation, the second defining the relation's attributes alongside their respective labels and the third defining the relations data followed for the values of each attributes for each record. The dataset collected was divided into two parts: training and testing data – the training data was used to formulate the model while the test data was used to validate the model. The process of training and testing predictive model according to literature is a very difficult experience especially with the various available validation procedures.

e Model validation and evaluation metrics

For this classification problem, it was natural to measure a classifier's performance in terms of the error rate. The classifier predicted the class of each instance – the record containing values for each risk of hypertension: if it is correct, that is counted as a success; if not, it is an error. The error rate being the proportion of errors made over a whole set of instances, and thus measured the overall performance of the classifier. The error rate on the training data set was not likely to be a good indicator of future performance; because the DT classifiers were been learned from the very same training data. In order to predict the performance of a classifier on new data, there was the need to assess the error rate of the predictive model on a dataset that played no part in the formation of the classifier. This independent dataset was called the test dataset – which was a representative sample of the underlying problem as was the training data.

The process of leaving a part of a whole dataset as testing data while the rest is used for training the model is called the holdout method. The challenge here is the need to be able to find a good classifier by using as much of the whole historical data as possible for training; to obtain a good error estimate and use as much as possible for model testing. It is a common trend to holdout one-third of the whole historical

dataset for testing and the remaining two-thirds for training. For this study the cross-validation procedure was employed, which involved dividing the whole datasets into a number of folds (or partitions) of the data. Each partition was selected for testing with the remaining $k - 1$ partitions used for training; the next partition was used for testing with the remaining $k - 1$ partitions (including the first partition used or testing) used for training until all k partitions had been selected for testing. The error rate recorded from each process was added up with the mean the mean error rate recorded. The process used in this study was the stratified 10-fold cross validation method which involves splitting the whole dataset into ten partitions. Figure 2 shows a representation of the 10-fold cross validation process.

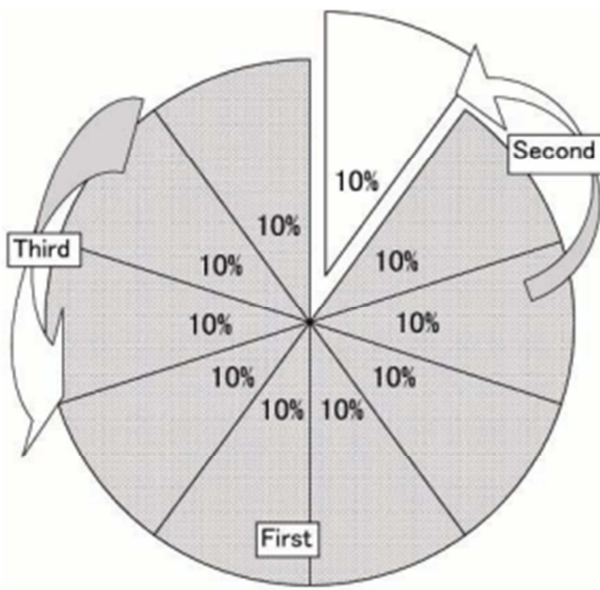


Figure 2. 10-fold cross validation process.

During the course of evaluating the predictive model, a number of metrics were used to quantify the model's performance. In order to determine these metrics, four parameters 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 (FN). In this study which involves a binary classification, either of survived and not survived can be considered as positive while the others negative (e.g. if low is considered as a positive then moderate and high are negatives and vice versa).

True positives are the correct prediction of positive cases, true negatives are the correct prediction of negative cases, false positives are the negative cases predicted as positives while false negatives are positive cases predicted as negatives. These results are presented on confusion matrix (Figure 3) – for this study the confusion matrix is a 3 x 3 owing to the three labels for the output class but for simplicity of the notion of positives and negatives using a 2 x 2 confusion matrix is presented. The four parameters were used to formulate the metrics discussed as follows:

The performance metrics are thus defined as follows:

- *Sensitivity/True positive rate/Recall*: is the proportion of actual positive cases that were correctly predicted positive by the model.

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad (7)$$

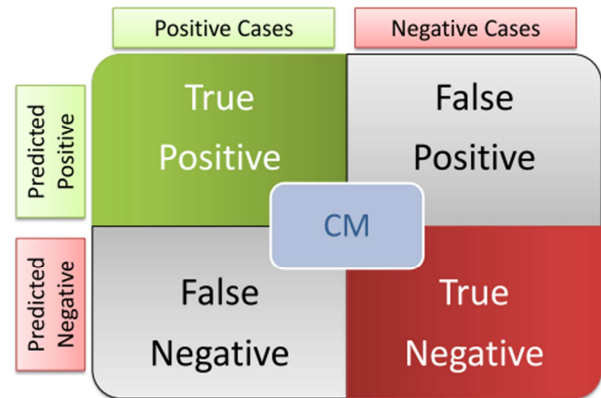


Figure 3. Confusion matrix for model performance.

- *False Positive rate/False alarm*: is the proportion of actual negative cases that were predicted as positive by the model.

$$\text{False alarm} = \frac{FP}{FP+TN} \quad (8)$$

- *Precision*: is the proportion of the predicted positive/negative cases that were actually positive or negative. Equations (9) and (10) show the precision for positive and negative cases.

$$\text{Precision (positive class)} = \frac{TP}{TP+FP} \quad (9)$$

$$\text{Precision (negative class)} = \frac{TN}{FN+TN} \quad (10)$$

- *Area under the Receiver Operating Characteristics (ROC) curve*: is the area of the curve plotted by the graph of the true positive rate (sensitivity) against the true negative rate (specificity) for the different instances of test datasets used for testing the predictive model for hypertension risk.
- *Accuracy*: is the total number of correct classifications (positive and negative)

$$\text{Accuracy} = \frac{TN+TP}{TN+TP+FN+FP} \quad (11)$$

4. Results and Discussion of Predictive Model for Hypertension Risk

For this study, data was collected from 32 patients using the questionnaires constructed for this study among which; the risk of hypertension was identified for 22 patents while the risk of hypertension was not identified for the remaining 10 patients. This was done in order to use the dataset of the 22 patients to formulate (train and test) the predictive model

for the risk of hypertension while the dataset of the remaining 10 patients was used for the external validation of the predictive model in order to determine the risk of hypertension using the formulated and validated model.

Figure 4 shows a screenshot of the data collected from the 32 respondents selected for this study. The data was stored in the attribute relation file format (.arff) which is the acceptable format for the data mining simulation environment selected for this study. The format required the identification of three (3) parts of the dataset, namely:

- a *The relation section:* was used to identify the name of the file identified which in this case is hypertension-all for the data containing all 32 patients, hypertension-training for the data containing the 22 respondents selected for training and testing the model after formulation and hypertension-testing for the data containing the 10 respondents which were used for the external validation of the model. The relations tag is identified using the name @relation before the relation name;

- b *The attribute section:* was used to identify the fields/attributes (risk factors) identified as the input variables for the risk of hypertension where the last attributes describes the risk of hypertension. There are 22 attributes identified in the file with the first 21 identifying the input variables (risk factors of the risk of hypertension) while the last variable is the risk of hypertension. Each attributes has its own respective label which shows the possible values that can be stated by each attribute defined in the dataset. The attribute tag for each attribute is identified using the name @attribute before each attribute name; and
- c *The data section:* was used to identify the dataset values for each respondents collected in the same order as the attributes were listed. Each respondent's record of data is represented as the set of values on each line with the risk of hypertension shown on the last portion of each line. The data containing the values of the attributes for each respondent is listed on the line following the name tag identified as @data.

```

1 @relation hypertension-all
2
3 @attribute Sex {Male,Female}
4 @attribute Age {below-30yrs,31-40yrs,41-50yrs,51-60yrs,above-60yrs}
5 @attribute Education {None,FSLC,SSCE,NCE,OND,HND,BSc}
6 @attribute Occupation {Trader,Public-service,Private,Artisan,None}
7 @attribute Marital-Status {Single,Married}
8 @attribute Position {Operations,Manager,Teacher,Marketer,Student,Clerk,Farmer}
9 @attribute Residence {Village,Town,City}
10 @attribute Ethnicity {Ibo,Hausa,Yoruba,Delta}
11 @attribute Family-History {No,First,Second}
12 @attribute BMI {Normal,Underweight,Obese,Overweight}
13 @attribute Exercise {Daily,Weekly,Rarely,Never}
14 @attribute Marital-issues {Never,Sometimes,Always}
15 @attribute Stress-work {Never,Sometimes,Always}
16 @attribute length-work {1,2,3,4,5,6,7,8,9,10,11,12,13,14,15}
17 @attribute smoke {Yes,No}
18 @attribute smoke-amt {Nil,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15}
19 @attribute alcohol {No,Yes,Quit}
20 @attribute alc-brand {Nil,Stout,Beer,Gin,Wine}
21 @attribute alc-amt {Nil,1,2,3,4,5}
22 @attribute alc-length-yrs {Nil,1,3,5,6,8,15,20,35}
23 @attribute salt-intake {Low,Moderate,High}
24 @attribute risk {Low,Moderate,High}
25
26 @data
27 Female,31-40yrs,NCE,Public-service,Married,Teacher,Town,Yoruba,No,Normal,Daily,Never,Never,8,No,Nil,No,Nil,Nil,Nil,High,Moderate
28 Male,above-60yrs,SSCE,Public-service,Married,Clerk,Town,Yoruba,No,Underweight,Daily,Never,Sometimes,10,No,Nil,Quit,Stout,1,3,Moderate,Low
29 Female,above-60yrs,FSLC,Artisan,Married,Manager,City,Yoruba,First,Obese,Rarely,Never,Sometimes,8,No,Nil,Yes,Stout,1,1,High,Moderate
30 Male,above-60yrs,None,Artisan,Married,Operations,Town,Yoruba,No,Normal,Daily,Never,Never,8,No,Nil,No,Nil,Nil,Nil,Moderate,Low
31 Female,51-60yrs,FSLC,Trader,Married,Marketer,Town,Yoruba,No,Obese,Never,Always,Always,12,No,Nil,No,Nil,Nil,Nil,High,High

```

Figure 4. Data collected for all 32 respondents collected.

Table 1 gives a description of the number of patients with their respective risk of hypertension from the 22 patient records selected for model formulation and validation which were stored in the hypertension-training.arff file. The table shows that out of the 22 patients considered; 8 (36.4%) had low risk of hypertension, 4 (18.1%) had high risk of hypertension while 10 (45.5%) had moderate risk of hypertension. It was observed that the highest case presented was for respondents with moderate risk of hypertension while the least case was presented for respondents with high risk of hypertension. From the dataset that was collected, it was observed that majority were female – bout 71.9% of all respondents considered for the study with just 28.1% male.

Table 1. Distribution of hypertension risk among historical dataset.

| Hypertension risk | Frequency | Percentage (%) |
|-------------------|-----------|----------------|
| Low | 8 | 36.4 |
| Moderate | 10 | 45.5 |
| High | 4 | 18.1 |
| Total | 22 | 100.0 |

Majority of the respondents also lies in the age group of above 60 years (53.1%) and 25% of the respondents within the age group of 51 to 60 years of age while the remaining 21.9% belonged to the age group of below 50 years. Also, 31.25% of the respondents had educational qualifications of first School leaving certificate (primary school) and senior school certificate (SSCE) but 18.8% of respondents had no formal education. 50% of the respondents were traders by occupation, 21.88% were public servants while 15.6% were

artisans by profession. 93.75% of the respondents were married while the remaining percentage was single and 50% of the respondents worked as marketers with 21.88% working at the operations. Majority of the respondents live in the town (71.88%) with most of the respondents been Yoruba (90.6%).

Based on the clinical information inferred from the respondents, the following were identified from the information provided. 68.75% of respondents had no family history of hypertension with the remaining having first generation family members with hypertension. 53.1% of the respondents had a normal BMI while about 18.8% were obese. 53.13% of respondents had daily exercises with 18.75% having their exercise done weekly or rarely. 50% of the respondents had issues at home – either with their spouses or from the children while 25% sometimes experience issues at home. About 50% of the respondents never experience stress at the office but 28% sometimes experience stress at the office. From the respondents, it was observed that the minimum hours spent in the office was 4 hours while the maximum hours spent was 8 hours with an average of 7 hours spent at the office.

Based on the social habits of the respondents, the following information were identified. 29 (90.62%) of respondents do not engage in smoking but the remaining percentage owed for 3 smokers for which the minimum sticks consumed was 4 with a maximum consumption of 10 sticks daily. Out of the respondents, 12 (37.5%) indulged in the habit of drinking out of which 7 drink stout, 3 drink beer while 1 drink Gin and Wine; the maximum bottles consumed was 2 with a minimum of 1. Also, the maximum length of years of consuming alcohol was 35 years with a minimum length of 5 years.

a Model formulation and simulation results using the C4.5 DT algorithm

From the dataset collected from the respondents, the training data was used for the formulation of the predictive model needed for the prediction of the risk of hypertension. The J4.8 decision trees algorithm was used to implement the C4.5 decision trees algorithm for the formulation of the predictive model using the simulation environment. The results of the formulation of the predictive model for the risk of hypertension using the C4.5 decision trees algorithm showed that four (4) variables were the most important risk factors of hypertension and were used by the algorithm to develop the tree that was used in formulating the predictive model for risk of hypertension using the C4.5 decision trees algorithm. The variables identified in the order of their importance were:

- Alcohol brand;
- Exercise;
- Marital issues; and
- Gender of the respondent.

Based on the four (4) variables identified by the C4.5 decision trees algorithm, the predictive model for the risk of hypertension was formulated based on the results of the simulation using the J48 decision trees algorithm on the

WEKA simulation environment. Figure 5 shows the decision trees that was formulated based on the four (4) variables that were proposed by the algorithm. The tree was used to deduce the set of rules that were proposed for determining the risk of hypertension based on the values of the four variables identified by the algorithm.

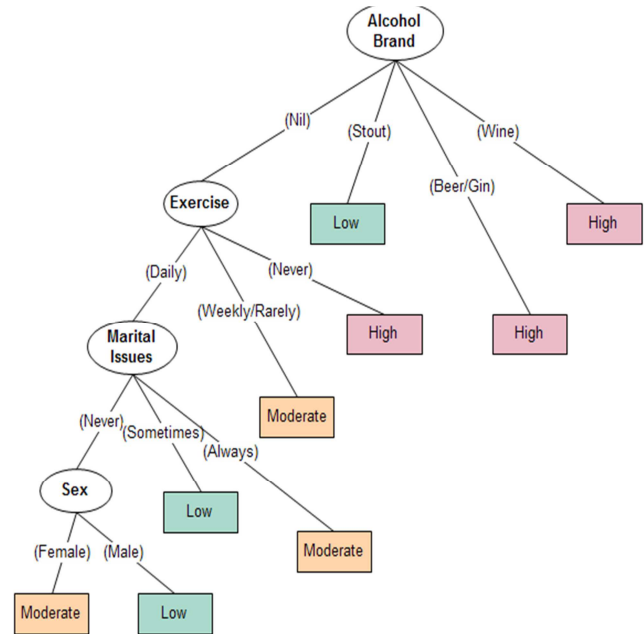


Figure 5. Decision Tree formulated using C4.5 for Risk of Hypertension.

The rules extracted from the tree are as follows:

- IF (alcohol brand = nil) AND (exercise = daily) AND (marital issues = never) and (sex = male) THEN (risk of hypertension = Low);
- IF (alcohol brand = nil) AND (exercise = daily) AND (sex = female) THEN (risk of hypertension = Moderate);
- IF (alcohol brand = nil) AND (exercise = daily) AND (marital issues = sometimes) THEN (risk of hypertension = Low);
- IF (alcohol brand = nil) AND (exercise = daily) AND (marital issues = rarely) THEN (risk of hypertension = Moderate);
- IF (alcohol brand = nil) AND (exercise = weekly) THEN (risk of hypertension = Moderate);
- IF (alcohol brand = nil) AND (exercise = rarely) THEN (risk of hypertension = Moderate);
- IF (alcohol brand = nil) AND (exercise = never) THEN (risk of hypertension = High);
- IF (alcohol brand = stout) THEN (risk of hypertension = Low);
- IF (alcohol brand = beer) THEN (risk of hypertension = High);
- IF (alcohol brand = gin) THEN (risk of hypertension = Moderate); and
- IF (alcohol brand = wine) THEN (risk of hypertension = High).

Following the simulation of the predictive model for risk

of hypertension using the C4.5 decision trees algorithm, the evaluation of the performance of the model following internal validation using the 10-fold cross validation method was recorded. Figure 6 shows the confusion matrix that was used to interpret the results of the true positive and negative alongside the false positive and negatives of the validation results. The confusion matrix shown was used to evaluate the performance of the predictive model for risk of hypertension.

| A | B | C | ← Classified as |
|---|---|---|--------------------------|
| 7 | 1 | 0 | A = Low Risk |
| 2 | 8 | 0 | B = Moderate Risk |
| 0 | 0 | 4 | C = High Risk |

Figure 6. Confusion matrix of performance evaluation using C4.5.

From the confusion matrix shown in figure 4.2, the following sections present the results of the model's performance. Out of the 8 low cases, there were 7 correct classifications with 1 misclassified as moderate risk; out of the 10 moderate risk cases, there were 8 correct classifications with 2 misclassified as low risks and all 4 high risk cases were correctly classified. Therefore, there were 19 correct classifications out of the 22 records considered for the model development owing for an accuracy of 86.36%. Table 2 shows the summary of the evaluation results.

Table 2. Summary of the results of performance evaluation using C4.5.

| Performance Metrics | Risk Labels | Values | Average |
|------------------------------|-------------|--------|---------|
| TP rate (sensitivity/recall) | Low | 0.875 | 0.892 |
| | Moderate | 0.800 | |
| | High | 1.000 | |
| FP rate (false alarm rate) | Low | 0.143 | 0.075 |
| | Moderate | 0.083 | |
| | High | 0.000 | |
| Precision | Low | 0.778 | 0.889 |
| | Moderate | 0.889 | |
| | High | 1.000 | |
| ROC | Low | 0.929 | 0.954 |
| | Moderate | 0.933 | |
| | High | 1.000 | |

b Model formulation and simulation results using the ID3 DT algorithm

From the dataset collected from the respondents, the training data was used for the formulation of the predictive model needed for the prediction of the risk of hypertension. The ID3 decision trees algorithm was used to implement the ID3 decision trees algorithm for the formulation of the predictive model using the simulation environment. The results of the formulation of the predictive model for the risk of hypertension using the ID3 decision trees algorithm

showed that four (4) variables were the most important risk factors of hypertension and were used by the algorithm to develop the tree that was used in formulating the predictive model for risk of hypertension using the ID3 decision trees algorithm. The variables identified in the order of their importance were:

- Length of work (in hours);
- Marital issues;
- Sex; and
- Occupation.

Based on the four (4) variables identified by the ID3 decision trees algorithm, the predictive model for the risk of hypertension was formulated based on the results of the simulation using the J48 decision trees algorithm on the WEKA simulation environment. Figure 7 shows the decision trees that was formulated based on the four (4) variables that were proposed by the algorithm. The tree was used to deduce the set of rules that were proposed for determining the risk of hypertension based on the values of the four variables identified by the algorithm.

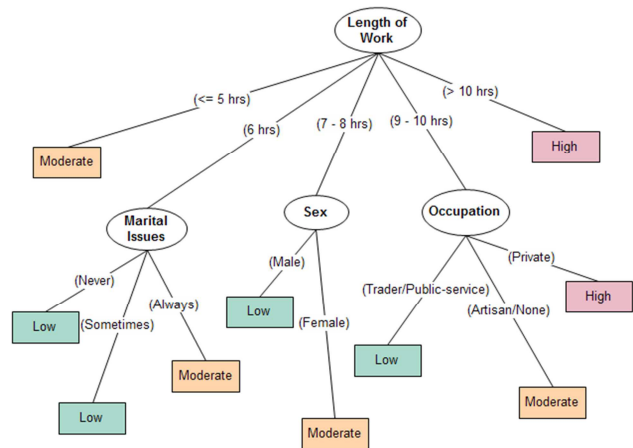


Figure 7. Decision Tree formulated using ID3 for Risk of hypertension.

The rules extracted from the tree are as follows:

- IF (length of work ≤ 5 hours) THEN (risk of hypertension = Moderate);
- IF (length of work = 6 hours) AND (marital issues = never/sometimes) THEN (risk of hypertension = Low);
- IF (length of work = 6 hours) AND (marital issues = always) THEN (risk of hypertension = Moderate);
- IF (length of work = 7-8 hours) AND (sex = male) THEN (risk of hypertension = Low);
- IF (length of work = 7-8 hours) AND (sex = female) THEN (risk of hypertension = Moderate);
- IF (length of work = 9-10 hours) AND (occupation = trader/public service) THEN (risk of hypertension = Low);
- IF (length of work = 9-10 hours) AND (occupation = artisan/none) THEN (risk of hypertension = Moderate);
- IF (length of work = 9-10 hours) AND (occupation = private) THEN (risk of hypertension = High); and

ix IF (length of work > 10 hours) THEN (risk of hypertension = High).

Following the simulation of the predictive model for risk of hypertension using the ID3 decision trees algorithm, the evaluation of the performance of the model following internal validation using the 10-fold cross validation method was recorded. Figure 8 shows the confusion matrix that was used to interpret the results of the true positive and negative alongside the false positive and negatives of the validation results. The confusion matrix was used to evaluate the performance of the predictive model for risk of hypertension.

| A | B | C | ← Classified as |
|---|----|---|--------------------------|
| 8 | 0 | 0 | A = Low Risk |
| 0 | 10 | 0 | B = Moderate Risk |
| 0 | 0 | 4 | C = High Risk |

Figure 8. Confusion matrix of performance evaluation using ID3.

From the confusion matrix shown in figure 8, the following sections present the results of the model's performance. Out of the 8 low cases, all were correctly classified; out of the 10 moderate risk cases, all were correctly classified and out of the 4 high risk cases, all were correctly classified. Therefore, the overall accuracy of the ID3 algorithm was 100%. Table 3 shows the summary of the evaluation results.

Table 3. Summary of the results of performance evaluation using ID3.

| Performance Metrics | Risk Labels | Values | Average |
|-------------------------------|-------------|--------|---------|
| TP rate (sensitivity/recall) | Low | 1.000 | 1.000 |
| | Moderate | 1.000 | |
| | High | 1.000 | |
| FP rate (false alarm rate) | Low | 0.000 | 0.000 |
| | Moderate | 0.000 | |
| | High | 0.000 | |
| Precision | Low | 1.000 | 1.000 |
| | Moderate | 1.000 | |
| | High | 1.000 | |
| ROC | Low | 1.000 | 1.000 |
| | Moderate | 1.000 | |
| | High | 1.000 | |

The result of the performance evaluation of the C4.5 algorithm was presented in Table 2. The true positive rate which gave a description of the proportion of actual cases that was correctly predicted showed values of 0.875, 0.8 and 1.00 for the low, moderate and high cases respectively presented owing for an average value of 0.892 – 89.2% of the actual risks was correctly predicted by the predictive model. The false positive rate which gave a description of the proportion of predicted cases that was incorrectly classified showed values of 0.143, 0.083 and 0.00 for the low, moderate and high

risk cases respectively presented owing for an average value of 0.075 – 7.5% of the predicted cases were misclassified. The precision which gave a description of the proportion of the predicted cases that was correctly classified showed values of 0.778, 0.889 and 1.00 for the low, moderate and high risk cases respectively owing for an average value of 0.889 – 88.9% of the predicted cases were actually correct. The receiver operating characteristics curve which gave a description of how well the predictive model was able to discriminate between all three risk cases showed values of 0.929, 0.933 and 1.00 for the low, moderate and high risk cases respectively owing for an average value of 0.954 – 95.5% of the area under the ROC was covered in the plot of the TP rate against the FP rate, which implies that the model showed good capability of discriminating between the risks of hypertension.

The results of the performance evaluation of the ID3 algorithm shown in Table 3 as presented is discussed in the following paragraphs. The true positive rate which gave a description of the proportion of actual cases that was correctly predicted showed values of 1.000 each for the low, moderate and high cases presented owing for an average value of 1.000 – 100% of the actual risks was correctly predicted by the predictive model. The false positive rate which gave a description of the proportion of predicted cases that was incorrectly classified showed values of 0.000 each for the low, moderate and high risk cases presented owing for an average value of 0.000 – 0% of the predicted cases were misclassified. The precision which gave a description of the proportion of the predicted cases that was correctly classified showed values of 1.000 each for the low, moderate and high risk cases owing for an average value of 1.000 – 100% of the predicted cases were actually correct. The receiver operating characteristics curve which gave a description of how well the predictive model was able to discriminate between all three risk cases showed values of 1.000 each for the low, moderate and high risk cases owing for an average value of 1.000 – 100% of the area under the ROC was covered in the plot of the TP rate against the FP rate, which implies that the model showed good capability of discriminating between the risk of hypertension.

Based on the results of the evaluation of the performance of the two (2) decision trees algorithm proposed for this study, the ID3 decision trees algorithm was observed to show the best performance based on the results of the true positive (TP) rate, false positive (FP) rate, precision, area under the ROC curve and the accuracy of the predictive model's validation using the training dataset for the study.

5. Conclusion

This study focused on the development of a prediction model using identified risk factors in order to classify the risk of hypertension in selected respondents for this study. Historical dataset on the distribution of the risk of hypertension among respondents was collected using questionnaires following the identification of associated risk factors of hypertension from expert medical practitioners. The dataset

containing information about the risk factors identified and collected from the respondents was used to formulate predictive model for the risk of hypertension using two (2) decision trees algorithm – C4.5 and ID3. The predictive model development using the decision trees algorithm was formulated and simulated using the WEKA software.

Following the identification of the features relevant for the risk of hypertension among the respondents selected for this study, four (4) variables were identified by both decision trees algorithm although two (2) of the variables were observed to be common among both models, namely: the marital issues and the sex of the respondents. The predictive model developed using the dataset showed good results with both algorithm having very high rates of properly discriminating between the different risks of hypertension with the results of the ID3 algorithm outperforming that of the C4.5 decision trees algorithm. The variables identified by the prediction model and the tree constructed from the decision tree using the variables can help provide insight into the relationship that exist between the variables (risk factors) with respect to the classification of the risk of hypertension.

Following the development of the prediction model for hypertension risk classification, a better understanding of the relationship between the attributes relevant to CML survival was proposed. The model can also be integrated into existing Health Information System (HIS) which captures and manages clinical information which can be fed to the hypertension risk classification prediction model thus improving the clinical decisions affecting hypertension risk and the real-time assessment of clinical information affecting hypertension risk from remote locations. It is advised that a continual assessment of variables monitored for hypertension risk be made in order to increase the number of information relevant to creating an improved prediction model for hypertension risk classification using the proposed model in this study.

References

- [1] Jing-Song, L., Hai-Yan, Y. and Xiao-Guang, Z. (2011). Data Mining in Hospital Information System. In New Fundamental Technologies in Data Mining, Prof. Kimito Funatsu (Ed.). ISBN: 978-953-307-547-1. Available from: <http://www.intechopen.com/books/new-fundamental-technologies-in-data-mining/data-miningin-hospital-information-system>.
- [2] Aqueel, A. and Shaikh, A. H. (2012). Data Mining Techniques to find out Heart Diseases: An Overview. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)* 1 (4): 1–6.
- [3] Chen, J., and Greiner, R. (1999). *Comparing Bayesian Network Classifiers*. In Proceedings of UAI-99: 101–108.
- [4] Agbelusi, O. (2014). Development of a predictive model for survival of HIV/AIDS patients in South-western Nigeria, Unpublished MPhil Thesis, Obafemi Awolowo University, Ile-Ife, Nigeria.
- [5] Idowu, P. A., Aladekomo, T. A., Williams, K. O. and Balogun, J. A. (2015). Predictive model for likelihood of Sick cell anaemia (SCA) among pediatric patients using fuzzy logic. *Transactions in networks and communications* 31 (1): 31–44.
- [6] Wajjee, A. K., Higgings, P. D. R. and Singal, A. G. (2013). A Primer on Predictive Models. *Clinical and Translational Gastroenterology* 4 (44): 1–4.
- [7] Lackland, D. T. and Weber, M. A. (2015). Global burden of cardiovascular disease and stroke: hypertension at the core. *The Canadian journal of cardiology* 31 (5): 569–71. doi:10.1016/j.cjca.2015.01.009.
- [8] Poulter, N. R., Prabhakaran, D. and Caulfield, M. (2015). Hypertension. *Lancet* 386(9995): 801–812. doi:10.1016/s0140-6736(14)61468-9.
- [9] [10Carretero, O. A. and Oparil, S. (2000). Essential hypertension, Part I: definition and etiology. *Circulation* 101 (3): 329–35. doi:10.1161/01.CIR.101.3.329.
- [10] Wokoma, F. S. and Alasia, D. D. (2011). Blood pressure pattern in Barako: a rural community in Rivers State, Nigeria. *Niger Health J* 11: 8-13.
- [11] World Health Organization, WHO (2011). *WHO Maps: Non-communicable disease trend in all countries*. World Health Global Report, World Health Organization.
- [12] Unachukwu, C. N., Agomuoh, D. I. and Alasia, D. D. (2008). Pattern of non-communicable diseases among medical admissions in Port Harcourt, Nigeria. *Niger J Clin Pract* 11: 14-17.
- [13] World Health Organization, WHO (2002). *The World Health Report: 2002: Reducing the Risks, Promoting Healthy Life*. World Health Organization, Geneva.
- [14] Omran, A. R. (1971). The epidemiologic transition: A theory of the epidemiology of population change. *Milbank Mem Fund Q* 49: 509-38.
- [15] Cressman, M. D. and Gifford, R. W. (1983). Hypertension and stroke. *Journal of the American College of Cardiology* 1: 521-527.
- [16] Post, W. S., Hill, M. N. and Dennison, J. L. (2003). High prevalence of target organ damage in young, African American Inner-city men with hypertension. *Journal of clinical hypertension* 5: 24-30.
- [17] Khakurel, S., Agrawal, R. K. and Hada, R. (2009). Pattern of end stage renal disease in a tertiary care center. *Journal of the Nepal Medical Association* 48: 126–130.
- [18] Cuspidi, C., Lonati, L., Sampieri, L., Michev, L., Macca, G., Rocanova, J. I., Salerno, M., Fusi, V., Leonetti, G. and Zanchetti, A. (2000). Prevalence of target organ damage in treated hypertensive patients: different impact of clinic and ambulatory blood pressure control. *Journal of Hypertension* 18: 803-809.
- [19] Neal, B., MacMahon, S. and Chapman, N. (2000). Effects of ACE inhibitors, calcium antagonists and other blood pressure lowering drugs: results of prospectively designed overviews of randomized controlled trials. *Lancet* 356: 1955-1964.
- [20] Nornchahal, K., Ashton, W. D. and Wood, D. A. (2000). Alcohol consumption, metabolic cardiovascular risk factors and hypertension in women. *Int J Epidemiol* 29: 57-64.

- [21] Laurenzi, M., Mancini, M., Menotti, A., Stamler, J., Stamler, R., Trevisan, M. and Zanchetti, A. (1990). Multiple risk factors in hypertension: results from the Gubbio study. *J Hypertens* 8 (Supplementary): 7-12.
- [22] Chukwuonye, I. I., Chuku, A., Onyeonoro, U. U., Madukwe, O. O., Oviasu, E. and Ogah, O. S. (2013). A rural and urban cross-sectional study on alcohol consumption among adult Nigerians in Abia state. *Int J Med Biomed Res* 2: 179-185.
- [23] Nwaneli, C. U. and Omejua, E. G. (2010). Clinical presentation and aetiology of hypertension in young adults in Nnewi South East Nigeria. *Afrimed J I*: 24-26.
- [24] Oladapo, O. O., Salako, L., Sadiq, L., Shoyinka, K., Adedapo, K. and Falase, A. O. (2012). Target-organ damage and cardiovascular complications in hypertensive Nigerian Yoruba adults: a cross-sectional study. *Cardiovasc J Afr* 23: 379-384.
- [25] Diwe, K. C., Enwere, O. O., Uwakwe, K. A., Duru, C. B. and Chineke, H. N. (2015). Prevalence and awareness of hypertension and associated risk factors among bank workers in Owerri, Nigeria. *Int J Med Biomed Res* 4 (3): 142-148.
- [26] Cheng, T. H., Wei, C. P. and Tseng, V. S. (2006). Feature Selection for Medical Data Mining: Comparisons of Expert Judgment and Automatic Approaches. In *Proceedings of the 19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06)*.
- [27] Shillabeer, A. and Roddick, J. (2007). Establishing a Lineage for Medical Knowledge Discovery. *ACM International Conference Proceeding Series* 311 (70): 29-37.
- [28] Cao, X., Maloney, K. B. and Brusic, V. (2008). Data mining of cancer vaccine trials: a bird's-eye view. *Immunome Research* 4:7 - 11. DOI:10.1186/1745-7580-4-7.
- [29] Mayilvaganan, M and Rajeswari, K. (2014). Human Blood Pressure Classification Analysis Using Fuzzy Logic Control System in Data Mining. *International Journal of Emerging Trends and Technology in Computer Science (IJETICS)* 3 (1): 305-306.
- [30] Waghmare, K. (2013). Prediction of Occurrence of Non-Communicable Diseases Depending on Lifestyle Habits Clustering Data Mining Technique. In *Proceedings of International Conference on Emerging Research in Computing, Information, Communication and Applications ERCICA, 2013*: 1-5.