Digital Innovations & Contemporary Research in SCIENCE & ENGINEERING
Vol. 5 No. 2, June 2017

Article Citation Format
F.A. Oladeji, Idowu, P.A., O. Komolafe & O. Oyetunji (2017). Predictive Model for the Risk of Hyperopia in Nigeria Using Decision Tree Algorithm. Journal of Digital Innovations & Contemp Res. In Sc., Eng & Tech Vol. 5, No. 2. Pp 79-100

Article Progress Time Stamps

Article Type: Research Article
Manuscript Received: 23rd April, 2017
Review Type: Blind
Review/Acceptance Information Sent : 10th June, 2017
Final Acceptance:: 11rd ' June, 2017
DOI Prefix: 10.22624
Series ISSN - 2488-8699

Predictive Model for the Risk of Hyperopia in Nigeria Using Decision Tree Algorithm

[1]F.A. Oladeji, [2]Idowu, P.A., [3]O. Komolafe & [4]O. Oyetunji
[1]Department of Computer Science, University of Lagos, Nigeria
[2]Department of Computer Science and Engineering, Obafemi Awolowo University, Ile-Ife, Nigeria
[3]Engineering Materials Development Institute, Federal Ministry of Science & Technology, Akure, Nigeria
[4]Department of Computer Science and Mathematics, Mountain Top University, Ibafo, Nigeria
[1]faoladeji"unilag.edu.ng, [2]paidowu@oauife.edu.ng, [4]mooyetunji@mtu.edu.ng

ABSTRACT

This paper presents a model to forecast the risk of hyperopia among Nigerians using decision trees algorithms based on historical information elicited about the risk of hyperopia among selected respondents in southwestern Nigeria. The knowledge of hyperopia and the factors related to its risk, a number of variables (risk factors) were identified. The identified risk factors of hyperopia were validated by an expert physician with more than 10 years' experience in medical practice before the instrument of data collection was constructed alongside the identification of respondents. The predictive model was formulated using two (2) decision trees algorithms namely: C4.5 Decision trees and Classification and Regression Trees algorithms 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 of the evaluation of the performance of the two (2) decision trees algorithm proposed for this paper, the CART decision trees performed better than the C4.5 decision trees algorithm. It was also able to make 12.8% (6 records) more correct classifications than the C4.5 decision trees algorithm. Unlike the C4.5 decision trees algorithm, the CART algorithm was able to produce a very low FP rate for the mild cases with a difference of 32.9%. The precision and area under the ROC graphs of the CART algorithm also showed better ability to predict high proportion of classes and capacity to discriminate classes effectively.

Keywords: Hyperopia Risk Factors, CART, C4.5, Prediction Model, Decision Trees

1. INTRODUCTION

Hyperopia occurs when one eye turns upward – a form of eye misalignment also known as strabismus. With this disease the eye may turn up all the time or only part of the time. Thus, hyperopia can be constant or intermittent; when intermittent the eye moves only part of the time, stress or tiredness may be the cause of the hyperopia (Kabari and Nwachukwu, 2012). Most forms of strabismus affect children most often, adults may be diagnosed with a form of hyperopia that does not appear until later in life. The causes are related to muscles in both eyes that are not balanced and working together while in adults the common cause are nerve palsies, stroke, thyroid disease, trauma and neurological disorders.

Clinical Decision Support Systems (CDSS) provide clinicians, staff, patients, and other individuals with knowledge and person-specific information, intelligently filtered and presented at appropriate times, to enhance health and health care (Osheroff *et al.*, 2006). Medical errors have already become the universal matter of international society. In 1999, IOM (American Institute of Medicine) published a report "To err is Human" (Kohn *et al.*, 2000), that indicated: First, the quantity of medical errors is incredible, the medical errors had already became the fifth lethal; Second, the most of medical errors occurred by the human factor which could be avoid via the computer system. Improving the quality of healthcare, reducing medical errors, and guarantying the safety of patients are the most serious duty of the hospital.

The clinical guideline can enhance the security and quality of clinical diagnosis and treatment, its importance already obtained widespread approval (Miller and Kearney, 2004). In 1990, clinical practice guidelines were defined as "systematically developed statements to assist practitioner and patient decisions about appropriate health care for specific clinical circumstances" (Field and Lohr, 2005). The clinical decision support system (CDSS) is any piece of software that takes as input information about a clinical situation and that produces as output inferences that can assist practitioners in their decision making and that would be judged as "intelligent" by the program's users (Musen, 1997).

Data mining is involve discovering of meaningful and useful information in large data repositories (Idowu, 2017; Idowu *et al*, 2015). Data mining can discover valuable but hidden knowledge from databases especially those used in storing health-related information about diseases affecting patients (Bakpo and Kabari, 2011). 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 (Aqueel and Shaikh, 2012). 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 (Chen and Greiner, 1999). 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.

Hyperopia cases are on the increase among Nigerians affecting more children than adults thus posing a serious threat of blindness before children reach adulthood. In Nigeria, hyperopia risk is not pre-determined before the onset of the disease and most times individuals are already showing symptoms of the disease before by the time they reach the doctors for complaints. Developed nations all over have systems in place that aid the early detection of diseases but developing nations like Nigeria lack the availability of systems that possess such capability. There is a need for a model that can be used to determine the risk of hyperopia in Nigerians so as to prompt early detection. This paper presents a predictive model for the risk of having hyperopia in Nigeria using two decision tree algorithms.

2. RELATED WORKS

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 (Agbelusi, 2014; Idowu *et al.*, 2015). 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 (Waijee *et al*, 2013a). There are several important differences between traditional explanatory research and predictive research. Explanatory research typically applies statistical methods to test causal hypothesis using prior theoretical constructs. In contrast, predictive research applies statistical methods and/or machine learning techniques, without preconceived theoretical constructs, to predict future outcomes (e.g. predicting the risk of hospital readmission) (Breiman, 1984).

Although, predictive models may be used to provide insight into casualty of pathophysiology of the outcome, casualty is neither a primary aim nor a requirement for variable inclusion (Moons *et al.*, 2009). Non-causal predictive factors may be surrogates for other drivers of disease, with tumor markers as predictors of cancer progression or recurrence being the most common example. Unfortunately, a poor understanding of the differences in methodology between explanatory and predictive research has led to a wide variation in the methodological quality of prediction research (Hemingway *et al.*, 2009).

Umesh *et al.* (2016) performed a review of image processing and machine learning techniques for eye disease detection and classification. In the study, an expert system for the diagnosis of eye disease was presented using eye images. Following image acquisition, the image was segmented; normalized using Daugman's normalized model and the relevant features were extracted using circular symmetric filter method from the normalized image. Following this, an encoding procedure was used in matching the pre-processed images to their respective match as identified in the original dataset collected from the study location.

Kabari and Nwachukwu (2012) applied neural networks and decision trees for the detection of eye disease. The data set used for the training and testing of the system was collected from Linsolar Eye Clinic, Port Harcourt and Odadiki eye clinic, Port Harcourt all in Nigeria. The total data is 400 from which 320 samples (80%) are randomly chosen and used as training patterns and tested with 80 instances (20%) of the same data set. The data set consist of evenly distributed men and women. Samples also consider age randomly collected from 18 years to 70 years. Decision trees and artificial neural network were used to formulate the predictive model for eye disease risk. Using the hybrid model, a success rate of 92% was achieved. This infers that combination of neural networks and decision tree technique is an effective and efficient method for implementing diagnostic problem.

Imberman (2011) used decision trees to find patterns in an ophthalmology dataset. Using a learn and prune methodology, decision tree analysis of 354 accommodative esotropic patients led to the discovery of two conjunctive variables hat predicted deterioration in the initial year of treatment better than what was previously determined using standard statistical methods. The dataset collected consisted of 5, 073 records consisting of 54 eye related disease. The results showed that the model developed had specificity of 37% and sensitivity of 98%.

3. METHODS

The methodology consists of a sequence of methods/techniques which started with the identification of the variables predictive of hyperopia alongside the data collection method used in gathering the required data needed for model development. The historical data collected contained records of individuals consisting of their respective values for each identified variables (risk factors) as inputs alongside the target variable (risk of hyperopia) as the output variable.

The decision trees algorithm used in formulating the predictive model was proposed alongside the process of model development using the historical data for training and testing the predictive model for the likelihood of hyperopia. The simulation environment used for the development of the predictive model was also stated alongside the performance evaluation metrics used in validating the predictive model developed for risk of Hyperopia. Figure 3.1 shows a description of the process involved in the methodology design of this study used in the development of the predictive model for hyperopia.

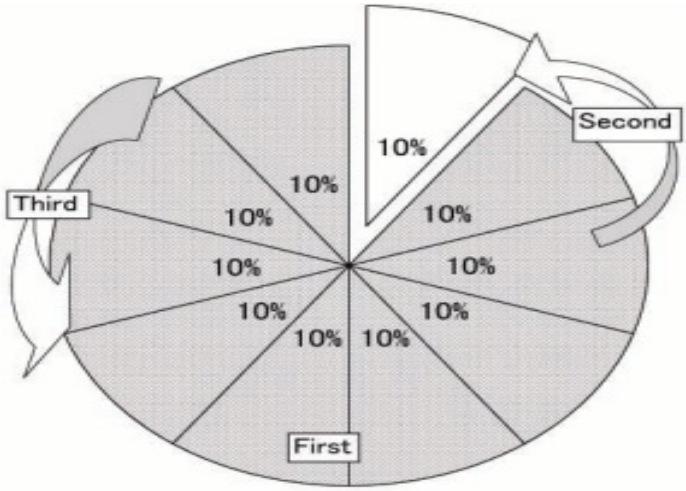

Figure 3.1: 10-fold cross validation process

3.1 Data Identification and Collection

Following the review of related works of literature in the body of knowledge of hyperopia and the factors related to its risk, a number of variables (risk factors) were identified. The identified risk factors of hyperopia were validated by an expert physician with more than 10 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 hyperopia but for those with the disease. Appendix I shows the questionnaire administered to the respondents selected for this study

3.1.1 Questionnaire design

Before the construction of the questionnaire, the expert physician provided information about the associated risk factors of hyperopia. The associated risk factors of hyperopia were classified as:

- **a.** Demographic – the demographic information used are: gender, age, marital status, ethnicity, occupation, religion and academic qualification; and
- **b.** Clinical factors – the clinical factors evaluated were axial length of the eye, family history of hyperopia, height (meters). Weight (Kg), curvature of the cornea, pathological conditions, presence of diabetes mellitus, vision blurred, use of contact lens, difficulty reading, and worn out lens.

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 hyperopia from the individual respondent.

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 hyperopia based on the information provided on each questionnaire. It is important to state that the comments provided by the physician is subjective to his own experience in medical practice and may not be a true representation of the generic risk of hypertension in Nigeria.

3.2 Formulation of the Predictive Model for Risk of Hyperopia

Following the identification and validation of variables relevant to the risk of hyperopia 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 hyperopia 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 hyperopia (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 hyperopia 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 hyperopia – 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 hyperopia) represented by the variable Y_j – the risk of hyperopia for the jth individual in the j records of data collected from the hospital selected for the study.

Equation 3.1 shows the mapping function that describes the relationship between the risk factors and the target class – risk of hyperopia.

$$\varphi: X \rightarrow Y \qquad (3.1)$$

$$defined\ as: \varphi(X) = Y$$

The equation shows the relationship between the set of risk factors represented by a vector, X consisting of the values of i risk factors and the label Y which defines the risk of hyperopia – low, moderate and high risk of hyperopia as expressed in equation 3.2. Assuming the values of the set of risk factors for an individual is represented as $X = \{X_1, X_2, X_3, \ldots\ldots, X_i\}$ where X_i is the value of each risk factor, i = 1 to i; then the mapping φ used to represent the predictive model for hyperopia risk maps the risk factors of each individual to their respective risk of hyperopia according to equation 3.2.

$$\varphi(X) = \begin{cases} Low\ risk \\ Mild\ risk \\ High\ risk \end{cases} \qquad (3.2)$$

The decision trees developed for the risk of hyperopia in individuals was used to propose a set of rules that can be used to determine the risk of hyperopia directly just by observing the values of the risk factors identified by the model and the succession of events. Also, the set of attributes identified in the final decision trees model for hyperopia are the risk factors which have the most relevant importance to the determination of the risk of hyperopia in individuals and was proposed to the physician to be given much consideration during hyperopia risk assessment of individuals. In the following section, the decision trees algorithm used in formulating the predictive model for hyperopia risk in individuals is presented. Although, it is important to stress that decision trees algorithm generates a hierarchical tree structure with a top-down structure using a splitting criteria with an underlying conditional probability using the succession of occurrence of risk factors.

Hence, the splitting criteria used by each decision trees algorithm presented was presented which is used to determine the selection of tree nodes starting from the parent node (the most important variable) to the successive nodes all the way to the leaf (target class representing the risk of hyperopia).

3.2.1 Decision tree (DT) classifier

The formulation of the predictive model for hyperopia risk in individuals was proposed using decision trees algorithm for the classification of the risk of hyperopia 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 hyperopia 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 hyperopia – 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.

3.2.2 Decision trees (DT) algorithms used

The training data set, S is a set containing $S_1, S_2, \ldots, S_j$ of already classified samples of the records of individuals consisting of the values of their risk factors, $X = \{X_1, X_2, \ldots, X_i\}$ alongside the risk of hyperopia, $Y = \{low, mild, high\}$ 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 hyperopia in individuals alongside their respective criteria were:

i. C4.5 Decision trees algorithm (implemented as J48 algorithm); and
ii. Classification and Regression Trees, CART (implemented as simple CART).

The theory of decision trees has the following parts: a root node which is the starting point of the trees with branches called edges connecting successive nodes showing the flow based on the values (edge for transition) of the attribute (node) and nodes that have child nodes are called interior nodes (parent nodes). Leaf or terminal nodes are those nodes that do not have child nodes and represent a possible value of the target variable (CML survival class) given the variables represented by the path from the root node. Rules can then be induced from the trees taking paths created from the root node all the way to their respective leaf using IF-THEN statements.

The basic idea of any decision trees analysis is to split the given dataset into subsets by recursive partitioning of the parent nodes into child nodes based on the homogeneity of the of within–node instances or separation of between-node instances with respect to their target variables. Thus at each nodes, attributes are examined and the splitter is chosen to be the attribute such that after dividing the nodes into child nodes according to the value of the attribute variable, the target is differentiated to the best using algorithm.

As a result of this, there is a need by the decision trees algorithm to distinguish between important variables attributes and attributes which contribute little to overall decision process which are based on the use of impurity measures. Following is the algorithm used by decision trees in growing their trees from a dataset containing a set of attributes. The algorithm is called *TreeGrowth* and takes in two arguments; which are the training records containing instances E and the attribute set (variables monitored) F which works by recursively splitting the data and expanding leaf nodes until a stopping criterion is met.

```
TreeGrowth(E, F)
If stopping_condition(E, F) = true then //test if the records have fallen below a threshold
        leaf = createNode( ) //create a leaf node if condition is met
        leaf.label = classify(E) //assign maximum CML 2-year survival class to leaf node
        Return leaf
else
        root = createNode( )//create root node if condition is not met
        root.test_condition = find_best_split(E, F) //determine attribute with the best split
        let V = {v| v is possible outcome of root.test_condition}//identify attribute splits
        for each v ϵ V do
                E_v = {e | root.test_condition(e) = v and e ϵ E}//assign each split to an edge
                child = TreeGrowth(E_v, F)//create a child tree at each edge
                add child as descendant of root and label the edge (root → child) as v.
                //child is the descendant tree along an edge (split) of root node (attribute)
        end for
end if
return root
```

The motivation for using the two (2) selected decision trees algorithms are as follows:

- They are one of the earliest classification models (Quinlan, 1986).
- They are very popular in medical data mining applications (Heath, D *et al.*, 1993; Murthy, 1998; Patel and Rana, 2014; Ahmad *et al.*, 2013).
- They are represented as an hierarchical tree structure consisting of attributes (as nodes) and attribute values as edges.
- They can be converted to If-Then rules.
- Attribute selection criteria uses a function that measures purity
  - C4.5 algorithm uses information-theoretic entropy.

3.3 Model Simulation Process and Environment

Following the identification of the decision tree (DT) learning algorithms that was needed for the formulation of the predictive model for the risk of hyperopia in pregnant women, the simulation of the predictive model was performed using the data collected which consisted of individuals records containing information about the risk factors and their respective risk of hyperopia collected from a hospital in south-western Nigeria. 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 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. 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 pregnant women's record containing values for each risk of hyperopia: 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. It was important that the test dataset was not used in any way to create the classifier since the machine learning classifiers involve two stages: one to come up with a basic structure of the predictive model and the second to optimize parameters involved in that structure.

3.3.1 10-fold cross validation technique

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 3.1 shows a representation of the 10-fold cross validation process.

For the purpose of this study, the explorer was used for performing the process of feature selection using 3 different feature selections each with its own unique search strategy. Following the identification of relevant indicators - input variables, the training dataset containing those instances were tested using the experimenter interface of the Weka environment. Thus, 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 hyperopia 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. Also, data can be read from comma separated values (.csv) format and from databases using Object-Database Connectivity (ODBC).

3.4 Performance Evaluation of Model 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 (FP). 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 – 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; a 2 x 2 confusion matrix is presented.

From the confusion matrix, it is easy to identify that the sum of the TP and the FP is the predicted positive cases while the sum of the TN and FN is the predicted negative cases. Also, the sum of the TP and the FN is the actual positive cases while the sum of the TN and the FP is the actual negative cases. 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.

$$Sensitivity = \frac{TP}{TP + FN} \quad (3.3)$$

- **Specificity/True negative rate:** is the proportion of actual negative cases that were correctly predicted as negatives by the model.

$$Specificity = \frac{TN}{FP + TN} \quad (3.4)$$

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

$$False\ alarm = \frac{FP}{FP + TN} \quad (3.5)$$

- **Precision:** is the proportion of the predicted positive/negative cases that were actually positive or negative. Equations (3.32) and (3.33) show the precision for positive and negative cases.

$$Precision\ (positive\ class) = \frac{TP}{TP + FP} \quad (3.6)$$

$$Precision\ (negative\ class) = \frac{TN}{FN + TN} \quad (3.7)$$

- **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)

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \quad (3.8)$$

4. RESULTS AND DISCUSSION

In this section, the results of the methodological approach described earlier are discussed. A thorough investigation into the analysis of the description of the dataset collected was initially performed in order to understand the distribution of the values of each risk factors of hyperopia among the respondents selected for this study using the minimum and maximum values, and the mean and standard deviation of the data distribution. Following this, the results of the model formulation and simulation process for the development of the predictive model for the risk of hyperopia was presented. The performance of the predictive models for hyperopia risk developed using the decision trees algorithms were evaluated in order to identify the most effective and efficient predictive model for the risk of hyperopia. Thus, the variables identified by the decision trees algorithm were proposed as the most important and relevant indicator for the risk of hyperopia among the patients selected for this study.

4.1 Results and Discussion of Data Summarization of Historical Dataset

For this study, data was collected from 50 patients using the questionnaires constructed for this study among which; the risk of hyperopia was identified for 47 patents. Figure 4.1 shows a screenshot of the data collected from the 47 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.

```
@relation Hyperopia-Risk-Model-Data

@attribute Gender {Male,Female}
@attribute Age {below-10,11-20,21-30,31-40,above-40}
@attribute Marital-status {Single,Married,Divorced}
@attribute Ethnicity {Yoruba,Ibo,Hausa,}
@attribute Occupation {Civil-servant,Trader,Teacher,Student}
@attribute Religion {Christian,Islam}
@attribute Qualification {Primary,Secondary,Tertiary,}
@attribute Axial-length-mm numeric
@attribute Family-history {None,First,Second}
@attribute Height-m numeric
@attribute Weight-kg numeric
@attribute Curvature numeric
@attribute Pathological-conditions {Yes,No}
@attribute Diabets-Mellitus {Yes,No,DK}
@attribute Visually-blurred {Yes,No,DK}
@attribute Contact-lens-use {Yes,No}
@attribute Difficulty-reading {Yes,No}
@attribute lens-worn-out {Yes,No}
@attribute Risk {Low,Mild,High}

@data
Female,21-30,Single,Yoruba,Civil-servant,Islam,Tertiary,22,First,1.3,67,39,No,No,Yes,Yes,Yes,Yes,Mild
Male,above-40,Divorced,Ibo,Teacher,Islam,Tertiary,22,None,1.4,61,39,No,No,Yes,No,No,Yes,Mild
Female,?,Divorced,Hausa,Civil-servant,Islam,Tertiary,24,First,1.8,82,40,No,DK,No,Yes,Yes,Yes,Low
Female,21-30,Married,Hausa,Civil-servant,Islam,Secondary,22,First,1.8,56,39,No,Yes,Yes,Yes,No,Yes,High
Female,21-30,Divorced,Hausa,Civil-servant,Islam,Tertiary,23,None,1.8,70,41,Yes,No,Yes,Yes,Yes,No,Mild
Female,21-30,Married,Hausa,Student,Islam,Tertiary,20,Second,1.5,67,39,Yes,Yes,Yes,Yes,Yes,No,Mild
Male,11-20,Single,Yoruba,Civil-servant,Islam,Secondary,21,First,1.5,81,40,Yes,No,Yes,No,Yes,No,Mild
Female,21-30,Married,Hausa,Teacher,Islam,Secondary,22,Second,1.63,67,39,Yes,No,?,No,Yes,No,Mild
Male,21-30,Single,Ibo,Civil-servant,Islam,Secondary,22,First,1.3,68,29,Yes,No,No,Yes,Yes,No,Mild
Male,11-20,Single,Yoruba,Trader,Christian,Secondary,23,Second,1.4,63,42,No,No,Yes,No,Yes,Yes,Mild
Female,11-20,Married,Hausa,Teacher,Christian,Secondary,21,Second,1.8,77,37,No,DK,No,No,No,Yes,Mild
Female,21-30,Single,Yoruba,Civil-servant,Islam,Secondary,21,None,1.4,61,40,No,Yes,Yes,Yes,No,Yes,High
Male,11-20,Married,Yoruba,Civil-servant,Islam,Secondary,23,None,1.5,93,34,Yes,Yes,Yes,No,No,No,High
Male,21-30,Married,Ibo,Teacher,Islam,Secondary,25,Second,1.4,69,38,Yes,No,Yes,No,Yes,No,Mild
Female,11-20,Married,Yoruba,Teacher,Islam,Secondary,22,None,1.6,77,41,Yes,Yes,Yes,No,Yes,No,Mild
Female,21-30,Single,Ibo,Civil-servant,Islam,Secondary,23,Second,1.6,83,39,Yes,No,Yes,No,?,Yes,Low
Male,below-10,Single,Yoruba,Student,Islam,Secondary,26,First,1.7,73,34,Yes,No,Yes,No,Yes,Yes,High
Female,11-20,Married,Ibo,Trader,Christian,Tertiary,26,None,1.7,77,39,Yes,No,Yes,No,No,Yes,High
```

Figure 4.1: Screenshot of the dataset collected from the respondents

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 hyperopia-model-training for the data containing all 47 patients selected for training and testing the model after. 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 hyperopia where the last attributes describes the risk of hypertension. There are 18 attributes identified in the file with the first 17 identifying the input variables (risk factors of the risk of hyperopia) while the last variable is the risk of hyperopia. Each attribute 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 hyperopia 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.

Table 4.1 gives a description of the number of patients with their respective risk of hyperopia from the 47 patient records selected for model formulation and validation which were stored in the hyperopia-training.arff file.

Table 4.1: Distribution of hyperopia risk among historical dataset

Hypertension risk	Frequency	Percentage (%)
Low	8	17.02
Mild	23	48.93
High	16	34.04
Total	47	100.0

The table shows that out of the 47 patients considered; 8 (17.02%) had low risk of hyperopia, 16 (34.04%) had high risk of hyperopia while 23 (48.93%) had mild risk of hyperopia. It was observed that the highest case presented was for respondents with mild risk of hyperopia while the least case was presented for respondents with high risk of hyperopia. Table 4.2 gives a description of the data collected from all 47 respondents selected for the study; it shows the distribution of the values of each attributes defined for the dataset collected from the respondents. From the summary of the dataset that was presented, it was observed that majority were female – with a ratio of 1.5:1 for female to men. The description also showed that 51.1% fell within the age interval of 21 – 30 years while 29.8% in the interval of 11 – 20; this showed that the majority of respondents selected for the study was between the age interval of 11 – 30 years with just 2.1% above 40 years. Furthermore, majority of the respondents were identified as Yoruba and Hausa with a proportion of 40.4% each with 48.9% respondents working as civil servants with the remaining proportion working as students, traders and teachers. The majority educational qualification was secondary school education with a proportion of 55.3% followed by tertiary education with a proportion of 29.8%.

Based on information regarding the clinical variables (risk factors of hyperopia), it was discovered that majority (40.4%) had first generation family members history of hyperopia followed by those with second generation (31.9%). 78.7% of the respondents had pathological conditions with about 25.53% of the respondents with diabetes mellitus while 21.3% don't know. 74.5% of respondents have had signs of visual blurriness while 46.8% of the respondents use contact lens and 72.3% had difficulty reading.

Table 4.2: Description of Historical Data of all 47 respondents

Variable Name	Labels	Frequency (%)
Gender	Male	19 (40.4%)
	Female	26 (59.6%)
Age (years)	Below 10	3 (6.38%)
	11-20	14 (29.78%)
	21-30	24 (51.06%)
	31-40	4 (8.51%)
	Above 40	1 (2.12%)
Marital Status	Single	20 (42.55%)
	Married	21 (44.68%)
	Divorced	6 (12.77%)
Ethnicity	Yoruba	19 (40.43%)
	Ibo	9 (16.14%)
	Hausa	19 (40.43%)
Occupation	Civil servant	23 (48.94%)
	Trader	8 (17.02%)
	Teacher	9 (19.15%)
	Student	7 (14.89%)
Religion	Christian	17 (36.17%)
	Islam	30 (63.83%)
Educational Qualification	Primary	7 (14.89%)
	Secondary	26 (55.32%)
	Tertiary	14 (29.77%)
Family History	None	12 (25.53%)
	First	19 (40.43%)
	Second	15 (31.91%)
Pathological Conditions	Yes	37 (78.72%)
	No	10 (21.28%)
Diabetes Mellitus	Yes	12 (25.53%)
	No	25 (53.19%)
	Don't Know	10 (21.28%)
Visually Blurred	Yes	35 (74.47%)
	No	10 (21.28%)
	Don't Know	1 (4.25%)
Contact Lens Use	Yes	22 (46.81%)
	No	25 (53.19%)
Difficulty reading	Yes	34 (72.34%)
	No	12 (25.53%)
Lens worn out	Yes	31 (68.09%)
	No	16 (31.91%)

4.2 Results of Model Formulation and Simulation

Following the identification of the risk factors that are associated with hyperopia risk, the next phase is model formulation using the aforementioned decision trees algorithms available in the Weka software. The 10-fold cross validation technique was used in evaluating the performance of the developed predictive model for hyperopia risk using the historical dataset used for training the model. This process was performed for both decision trees algorithm used with their respective performance compared for the most effective.

4.2.1 Results of model formulation and simulation using the C4.5 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 hyperopia using the C4.5 decision trees algorithm showed that four (4) variables were the most important risk factors of hyperopia 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:

a. Height (in metres);
b. Family History;
c. Marital Status; and
d. Occupation

Based on the four (4) variables identified by the C4.5 decision tees algorithm, the predictive model for the risk of hyperopia was formulated based on the results of the simulation using the J48 decision trees algorithm on the WEKA simulation environment. Figure 4.2 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 hyperopia based on the values of the four variables identified by the algorithm. The rules extracted from the tree are as follows:

Height-m <= 1.5
| Family-history = None: **High**
| Family-history = First: **Mild**
| Family-history = Second: **Mild**
Height-m > 1.5
| Marital-status = Single: **Low**
| Marital-status = Married
| | Occupation = Civil-servant: **Mild**
| | Occupation = Trader: **High**
| | Occupation = Teacher: **Mild**
| | Occupation = Student: **Mild**
| Marital-status = Divorced: **Low**

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 validation using the 10-fold cross validation method was recorded. Figure 4.3 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 in figure 4.3 was used to evaluate the performance of the predictive model for risk of hyperopia.

Digital Innovations
& Contemporary Research in
SCIENCE & ENGINEERING
Vol. 5 No. 2, June 2017

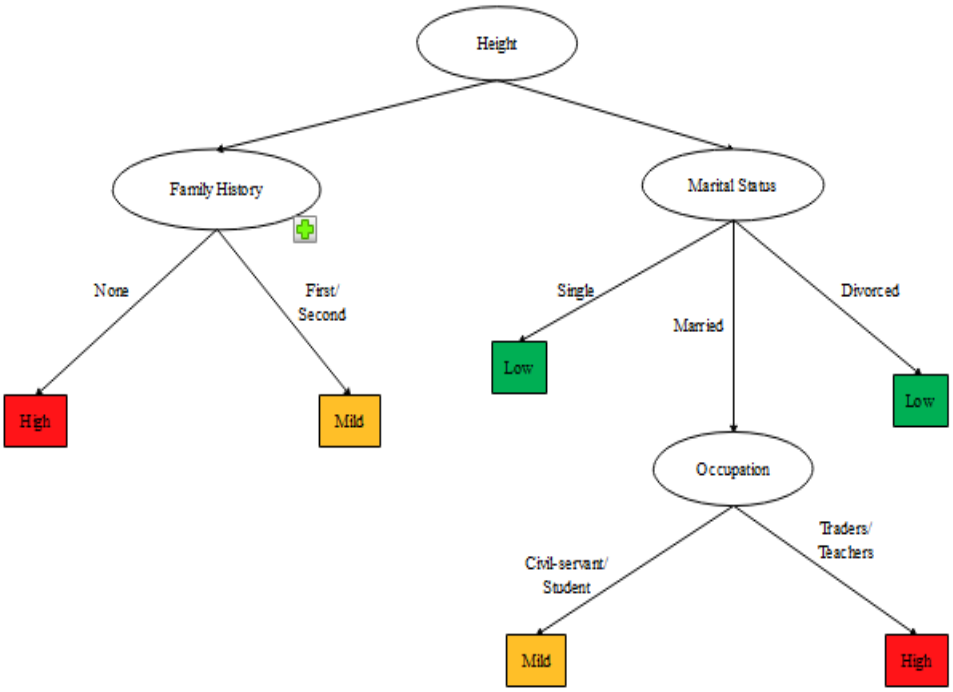

Figure 4.2: Decision Tree formulated using C4.5 for Risk of Hyperopia

a	b	c	<- - Predicted as
7	1	0	a = Low
1	21	1	b = Mild
2	7	7	c = High

Figure 4.3: Confusion matrix of performance evaluation using C4.5

From the confusion matrix shown in figure 4.3, 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 mild risk; out of the 23 mild risk cases, there were 21 correct classifications with 1 misclassified as low and high risks each while out of the 16 high cases, there were 7 correct classifications with 2 misclassified as low and 7 misclassified as high risk. Therefore, there were 35 correct classifications out of the 47 records considered for the model development owing for an accuracy of 74.47%. Table 4.3 shows the summary of the evaluation results.

4.2.2 Results of model formulation and simulation using the CART 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 hyperopia. The CART decision trees algorithm was used to implement the CART 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 hyperopia using the CART decision trees algorithm showed that eight (8) variables were the most important risk factors of hyperopia and were used by the algorithm to develop the tree that was used in formulating the predictive model for risk of hyperopia using the CART decision trees algorithm. The variables identified in the order of their importance were:

a. Curvature;
b. Weight (in Kilograms);
c. Difficulty reading;
d. Occupation;
e. Age;
f. Educational Qualification;
g. Height (in metres);
h. Family History; and
i. Ethnicity.

Table 4.3: Summary of the results of performance evaluation using C4.5

Performance Metrics	Risk Labels	Values
TP rate (sensitivity/recall)	Low	0.875
	Mild	0.913
	High	0.438
FP rate (false alarm rate)	Low	0.077
	Mild	0.333
	High	0.032
Precision	Low	0.700
	Mild	0.724
	High	0.875
ROC	Low	0.949
	Mild	0.855
	High	0.792

Based on the eight (8) variables identified by the CART decision tees algorithm, the predictive model for the risk of hyperopia was formulated based on the results of the simulation using the J48 decision trees algorithm on the WEKA simulation environment. Figure 4.4 shows the decision trees that was formulated based on the eight (8) 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 hyperopia based on the values of the variables identified by the algorithm.
The rules extracted from the tree are as follows:

Curvature < 33.0: **Low**
Curvature >= 33.0
| Weight-kg < 71.0
| | Weight-kg < 48.5: **High**
| | Weight-kg >= 48.5
| | | Difficulty-reading=(No)
| | | | Age=(above-40)|(31-40): **Mild**
| | | | Age!=(above-40)|(31-40): **High**
| | | Difficulty-reading!=(No): **Mild**
| Weight-kg >= 71.0
| | Height-m < 1.75
| | | Occupation=(Student): **High**
| | | Occupation!=(Student)
| | | | Qualification=(Tertiary)|(Secondary)
| | | | | Height-m < 1.625
| | | | | | Family-history=(Second)|(First)
| | | | | | Family-history!=(Second)|(First): **High**
| | | | | Height-m >= 1.625: **High**
| | | | Qualification!=(Tertiary)|(Secondary): **High**
| | Height-m >= 1.75: **Low**

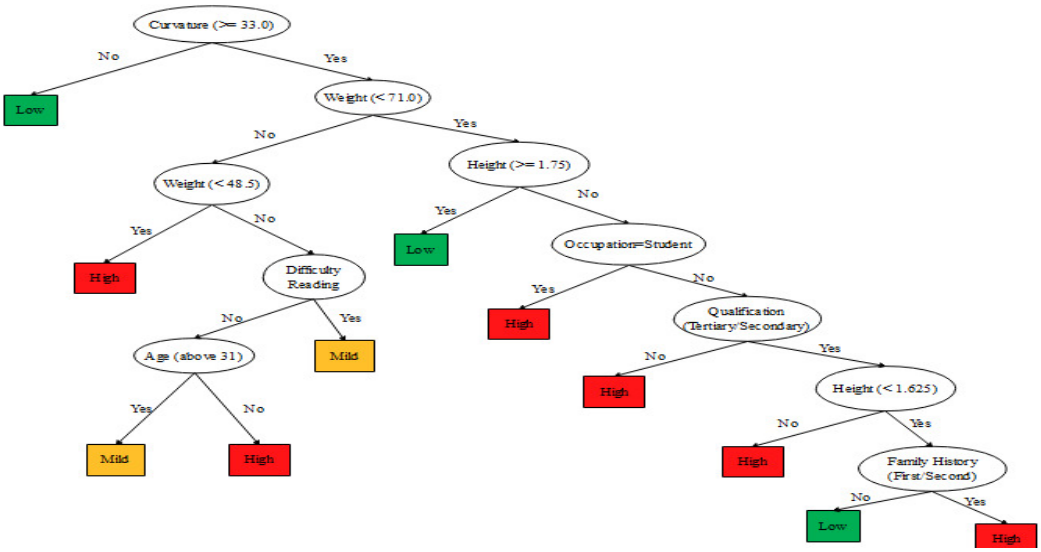

Figure 4.4: Decision Tree formulated using CART for Risk of Hyperopia

Following the simulation of the predictive model for risk of hyperopia using the CART decision trees algorithm, the evaluation of the performance of the model following internal validation using the 10-fold cross validation method was recorded. Figure 4.5 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 in figure 4.5 was used to evaluate the performance of the predictive model for risk of hyperopia. From the confusion matrix shown in figure 4.4, the following sections present the results of the model's performance. Out of the 8 low cases,7 were correctly classified while 1 was misclassified as mild risk; out of the 23 mild risk cases, 18 were correctly classified while 3 was classified as low and 2 misclassified as high risk and all 16 high risk cases were correctly classified. Therefore, the overall accuracy of the CART algorithm was 87.23%. Table 4.4 shows the summary of the evaluation results.

4.3 Discussion of results for model formulation and simulation

The result of the performance evaluation of the C4.5 algorithm was presented in Table 4.3. The true positive rate which gave a description of the proportion of actual cases that was correctly predicted showed values of 0.875, 0.913 and 0.438 for the low, mild and high cases respectively by the C4.5 decision trees algorithm.

a	b	c	<- - Predicted as
7	1	0	a = Low
3	18	2	b = Mild
0	0	16	c = High

Figure 4.5: Confusion matrix of performance evaluation using CART

Table 4.4: Summary of the results of performance evaluation using CART

Performance Metrics	Risk Labels	Values
TP rate (sensitivity/recall)	Low	0.875
	Mild	0.783
	High	1.000
FP rate (false alarm rate)	Low	0.077
	Mild	0.042
	High	0.065
Precision	Low	0.700
	Mild	0.947
	High	0.889
ROC	Low	0.949
	Mild	0.936
	High	0.991

The false positive rate which gave a description of the proportion of predicted cases that was incorrectly classified showed values of 0.077, 0.333 and 0.032 for the low, mild and high risk cases respectively. This implied that at most 8% and 33% of the low/high and mild risks respectively is misclassified as another class. The precision which gave a description of the proportion of the predicted cases that was correctly classified showed values of 0.700, 0.724 and 0.875 for the low, mild and high risk cases respectively. This implied that the algorithm was able to correctly classify at least 70% of class from a set of predictions.

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 0.949, 0.855 and 0.792 for the low, mild and high risk cases respectively. This implied that at least 79% 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 result of the performance evaluation of the CART algorithm was presented in Table 4.4. The true positive rate which gave a description of the proportion of actual cases that was correctly predicted showed values of 0.875, 0.783 and 1.00 for the low, mild and high cases respectively by the CART decision trees algorithm. The false positive rate which gave a description of the proportion of predicted cases that was incorrectly classified showed values of 0.077, 0.042 and 0.065 for the low, mild and high risk cases respectively. This implied that at most 8% of a class is misclassified as another class.

The precision which gave a description of the proportion of the predicted cases that was correctly classified showed values of 0.700, 0.947 and 0.889 for the low, mild and high risk cases respectively. This implied that the algorithm was able to correctly classify at least 70% of class from a set of predictions. 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 0.949, 0.936 and 0.991 for the low, mild and high risk cases respectively. This implied that at least 94% 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. Based on the results of the evaluation of the performance of the two (2) decision trees algorithm proposed for this paper, the CART decision trees performed better than the C4.5 decision trees algorithm. It was also able to make 12.8% (6 records) more correct classifications than the C4.5 decision trees algorithm. Unlike the C4.5 decision trees algorithm, the CART algorithm was able to produce a very low FP rate for the mild cases with a difference of 32.9%. The precision and area under the ROC graphs of the CART algorithm also showed better ability to predict high proportion of classes and capacity to discriminate classes effectively.

This paper presents the CART decision trees algorithm using the variables: curvature, weight (in Kilograms), difficulty reading, occupation, age, educational qualification, height (in metres), family history, and ethnicity to construct a decision trees for predicting the risk of hyperopia in Nigerian individuals

.

5.0 CONCLUSION

This paper focused on the development of a prediction model using identified risk factors in order to classify the risk of hyperopia in selected respondents for this study. Historical dataset on the distribution of the risk of hyperopia among respondents was collected using questionnaires following the identification of associated risk factors of hyperopia 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 hyperopia using two (2) decision trees algorithm – C4.5 and CART. The predictive model development using the decision trees algorithm was formulated and simulated using the WEKA software. The results of the paper revealed the variables that were identified by both decision trees algorithm as relevant for identifying the risk of hyperopia in respondents. The CART decision trees algorithm was observed to show a better accuracy compared to that of the C4.5 decision trees algorithm using the training dataset presented in the paper.

The paper proposed a predictive model for hyperopia risk using relevant risk factors selected from a predefined set of risk factors of hyperopia risk in Nigerians. The C4.5 decision trees algorithms identified four (4) risk factors while the CART decision trees algorithm identified eight (8) risk factors. The predictive model was formulated using the variables identified by both decision trees proposed for this paper and the performance evaluation of both models showed that the model developed using the CART decision trees algorithm was a better model. Unlike the C4.5 decision trees algorithm which selected nominal attributes (risk factors) alone, CART was able to identify nominal attributes in addition to the nominal values identified.

REFERENCES

1. Bakpo, F. S. and Kabari, L. G. (2011). Diagnosing Skin Diseases using an Artificial Neural Network. Artificial Neural Networks-Methodological Advances and Biomedical Applications. Suzuki, K. (Ed.) 978-9-53307-243-2, intech, Retrieved from http://www.intechopen.com/articles/show/title/diagnosing-skin-diseases-using-an-artificial-neural-network on June 23, 2016.
2. Banks, M.S. (1980). Infant refraction and accommodation. *Int Ophthalmol Clin 20*: 205 - 232.
3. Benjamin, W.J. and Borish, I.M. (2006). *Borish's clinical refraction* 2nd edition. St. Louis: Butterworth Heinemann: 9-11.
4. Chung, K.M., Mohidin, N. and Yeow, P.T. (1996). Prevalence of visual disorders in Chinese schoolchildren. *Optom Vis Sci 73*: 695 - 700.
5. Cook, R.C. and Glascock, R.E. (1951). Refractive and ocular findings in the newborn. *American Journal of Ophthalmology 34*: 1407 – 1413.
6. Crawford, H.E. and Haamar, G.E. (1949). Racial analysis of ocular deformities in schools of Hawaii. *Hawaii Med J 9*: 90 - 93.
7. Donders, F.C.(1964*). On the anomalies of accommodation and refraction of the eye.* London: New Syndenham Society: 80 - 86.
8. Duckman, R. (1990). Using photorefraction to evaluate refractive error, ocular alignment and accommodation in infants, toddlers and multiply handicapped children. *Probl Optom 2*: 333 - 353.
9. Field, M. J. and Lohr, K. N. (2005). *Clinical Practice Guidelines: Direction for a New Program*. Institute of Medicine, Committee on Clinical Practice Guidelines. Washington, DC. National Academy Press.
10. Gerali, P., Flom, M.C. and Raab, E.L. (1990). *Report of Children's Vision Screening Task Force.* Schaumburg, IL: National Society to Prevent Blindness.
11. Goldschmidt E. (1969). Refraction in the newborn. *Acta Ophthalmol 47*: 570 - 578.
12. Hammond, C.J., Snieder, H., Gilbert, C.E. and Spector, T.D. (2001). Genes and environment in refractive error: the twin eye study. *Invest Ophthalmol Vis Sci 42*: 1232 – 1261.
13. Hirsch, M.J. and Ditmas, D.L. (1947). Refraction of young myopes and their parents - a re-analysis. *Am J Optom 24*: 601 - 608.
14. Hirsch, M.J. and Weymouth, F.W. (1991). Prevalence of refractive anomalies. In: Grosvenor, T. and Flom, M. (eds.) *Refractive anomalies. Research and clinical applications*. Boston: Butterworth-Heinemann: 15-38.
15. Hsu-Winges, C., Hamer, R.D. and Norcia, A.M. (1988). Polaroid photo-refraction of infants: comparison to cycloplegic refraction. *J Pediatr Ophthalmol Strabismus 26*: 254 – 260.
16. Hung, L.F., Crawford, M.L. and Smith, E.L. (1995). Spectacle lenses alter eye growth and the refractive status of young monkeys. Nature Med 1: 761 - 765.
17. Idowu P. A. (2017). Predictive Model for the Classification of Hypertension Risk Using Decision Trees Algorithm. *American Journal of Mathematical and Computer Modelling*. 2(2): 48-59
18. Idowu, P. A., Aladekomo, T. A., Williams, K. O. and Balogun, J. A. (2015). Predictive model for likelihood of Sickle cell aneamia (SCA) among pediatric patients using fuzzy logic. *Transactions in networks and communications* 31 (1): 31–44.
19. Ingram RM, Walker C, Wilson JM, et al. Prediction of amblyopia and squint by means of refraction at age 1 year. *Br J Ophthalmol 70*: 12 - 15.

20. Ingram, R., Arnold, P., Dally, S. and Lucas, J. (1990). The results of a randomized trial of treating abnormal hypermetropia from the age of 6 months. *Br J Ophthalmol 74*: 158 - 159.
21. Ingram, R.M. and Walker, C. (1979). Refraction as a means of predicting squint or amblyopia in preschool siblings of children known to have these defects. *Br J Ophthalmol 63*: 238 - 242.
22. Ingram, R.M.(1977). Refraction as a basis for screening children for squint and amblyopia. *Br J Ophthalmol 61*: 8-15.
23. Jankiewicz, H. (1967). Embryologic and genetic factors in the refractive state of the eye. In: Hirsch, M.J. (ed.) Synopsis of the refractive state of the eye. *Am Acad Optom Ser 5*: 60 - 75. Minneapolis: Burgess Publishing.
24. Kabari, L.G. and Nwachukwu, E.O. (2012). Neural Networks and Decision Trees for Eye Diseases Diagnosis. *Advances in Expert Systems*: 63 – 84.
25. Kempf, G.A., Collins, S.D. and Jarman, E.L. (1928). Refractive errors in the eyes of children as determined by retinoscopic examination with a cycloplegic. *Public Health Bull no 182*. Washington DC: U.S. Government Printing Office.
26. Kohn, L. T., Corrigan, J. M. and Donaldson, M. S. (2000). *To err is human: building a safer health system*. Washington, D.C.: National Academy Press.
27. Miller, M. and Kearney, N. (2004). Guidelines for Clinical Practice: Development, Dissemination and Implementation. International Journal of Nursing Studies, 41 (7), 813 - 827.
28. Mohindra, I. and Held, R. (1981). Refraction in humans from birth to five years. *Doc Ophthalmol Proc Ser 28*: 19-27.
29. Morgan, M. (1958). Changes in refraction over a period of twenty years in a non-visually selected sample. *Am J Optom 35*: 281 – 299.
30. Musen, M. A. (1997). Modeling of Decision Support. *Handbook of medical informatics*. Bemmel, J.H.V. and Musen, M.A. (Eds.) Houten: Bohn Stafleu Van Loghum.
31. Osheroff, J.A., Teich, J.M. and Middleton, B.F. (2006). *A Roadmap for National Action on Clinical Decision Support*. American Medical Informatics Association. Retrieved from http://www.amia.org/inside/initiatives/cds/ on June 23, 2016.
32. Ottar, W.L., Scott, W.E and Holgado, S.I. (1995). Photoscreening for amblyogenic factors. *J Pediatr Ophthalmol Strabismus 32*: 289 – 295
33. Ottar, W.L., Scott, W.E. and Holgado, S.I. (1995). Photo-screening for amblyogenic factors. *J Pediatr Ophthalmol Strabismus 32*: 289 - 295.
34. Post, R.H. (1962). Population differences in visual acuity: review with speculative notes on selection relaxation. *Eugenics Quarterly 9*: 189 - 192.
35. Saunders, K.J. (1995). Early refractive development in humans. *Surv Ophthalmol 40*: 207 - 216.
36. Saunders, K.J., Woodhouse, M. and Westall, C.A. (1995). Emmetropisation in human infancy: rate of change is related to initial refractive error. *Vision Res 35*:1325-8.
37. Schmidt, P.P., Orel-Bixler, D. and Allen, D. (1995). Analytical comparisons of photorefraction in screenings. *Optom Vis Sci 72*: 209 – 217.
38. Simons, K. (1996). Preschool vision screening: rationale, methodology, and outcome. *Surv Ophthalmol 41*: 3 - 30.
39. Sorsby, A. (1972). *The functional anomalies Section I - Refraction and accommodation*. Modern ophthalmology, Philadelphia: JB Lippincott: 9-29.
40. Sorsby, A., Benjamin, B. and Davey, J.B. (1957). Emmetropia and its aberrations. *Medical Research Council Special Report* Series no. 293. London: Her Majesty's Stationery Office.

41. Sorsby, A., Benjamin, B., Sheridan, M. and Leary, G.A. (1961). Refraction and its components during growth of the eye from the age of three. *Medical Research Council Special Report* Series no. 301. London: Her Majesty's Stationery Office, 1961.
42. Sorsby, A., Leary, G.A. and Richards, M.J. (1962). Correlation ametropia and component ametropia. *Vision 2*: 309 - 313.
43. Tarczy-Hornoch, K. (2007). The epidemiology of early childhood hyperopia. *Optometry and Visual Science 84*: 115 - 123.
44. Umesh, L., Mrunalini, M. and Shinde, S. (2016). Review of Image Processing and Machine Learning Techniques for Eye Disease Detection and Classification. *International Research Journal of Engineering and Technology 3* (3): 547 – 551.
45. Van Alphen, G.W. (1961). On emmetropia and ametropia. *Ophthalmologica 142*: 1-9.
46. Williams, W.R., Latif, A.H.A., Hannington, L. and Watkins, D.R. (2005). Hyperopia and educational attainment in a primary school cohort. *Arch Dis Child 90* (2):150 - 153.