Model for Prediction of Cataracts Using Supervised Machine Learning Algorithms

Egejuru, N.C., Balogun, J.A., Mhambe, P.D., Asahiah, F.O. & Idowu, P.A.
Department of Computer Science & Engineering
Obafemi Awolowo University
Ile-Ife, Nigeria.

Corresponding author: paidowu@oauife.edu.ng

ABSTRACT

This study identified the risk factors for cataracts and formulated a predictive model based on the identified variables. The study simulated the formulated model and validated the model with a view to developing a model for cataracts' risk prediction. Following the review of the body of knowledge surrounding cataracts and their corresponding risk factors, interview with mental health professionals was conducted in order to validate the identified variables. Naïve Bayes, Decision Trees and the Multi-layer Perceptron classifiers were used to formulate the predictive model for the risk of cataracts based on the identified and validated variables using the WEKA software. The results of the data collected from 31 patients revealed 9 demographic variables and 17 risk factors variables alongside the respective risk factors, yielding a total of 26 variables in all. Out of the variables identified, the C4.5 Decision Trees algorithm revealed that smoking, myopia intensity, use of lenses and frequency of alcohol consumption were the most relevant risk factors out of cataract risks in the 26 variables identified. The results also showed that out of all the supervised machine learning algorithms used, the Multi-layer Perceptron was able to predict all records (100% accuracy) of the historical dataset used while the C4.5 Decision Trees and Naïve Bayes classifiers had an accuracy of 87% and 84% respectively. The study concluded that the Multi-layer Perceptron had the best capability to identify the unseen patterns existing within the variables used to formulate the predictive model for cataract's risks.

Keywords: Cataracts Risk Classification, Predictive Modeling, Machine Learning, Eye Disease

CISDI Journal Reference Format
Egejuru, N.C., Balogun, J.A., Mhambe, P.D., Asahiah, F.O. & Idowu, P.A. (2017) Model for Prediction of Cataracts Using Supervised Machine Learning Algorithms. Computing, Information Systems, Development Informatics & Allied Research Journal. Vol 8 No 3. Pp 47-62
Available online at www.cisdijournal.net

1. INTRODUCTION

The prevalence of visual impairment (VI) has been reported amongst different populations, with cataracts and refractive errors (RE) being reported as common causes. According to the World Health Organization [1], there are four levels of visual function, namely: normal vision; moderate visual impairment (VI); severe VI; and blindness. Moderate and severe VI are grouped together under the term 'low vision'; and low vision (LV) taken together with blindness represents total visual impairment. From a global perspective, uncorrected refractive errors are the main causes of moderate and severe visual impairment and cataract remains the leading cause of blindness in middle and low income countries [2].

Blindness is a global problem with socio-economic impact which does not only lead to the reduction of one's quality of life, but also reduces ones' capabilities and usefulness in the family, community and nation at large. The individual affected by blindness finds it difficult to have gainful employment, and this leads to dependence on family members and the society [3]. Seventy-five to eighty percent of blindness is avoidable [4] of which cataract is responsible for over half. Cataract is a major cause of avoidable blindness and visual impairment throughout the world and is likely to present an increasing burden to health care systems [5]. Hence, it is desirable to identify risk factors for the development and progression of cataract. Although surgical intervention is an effective modality for restoring vision, there are significant challenges in both delivery and utilization of cataract surgical services, especially among the most disadvantaged groups in the population [6].

Given the magnitude of blind people in Africa, Nigeria's population accounts for blind people estimated at 1,130,000 blind and aged over 40 years (4.2% prevalence of blindness) [7]. In a study by Isaac *et al*. [8], the majority of the world's 20 million cataract affected people live in the developing world with 6 million Africans who are blind as a consequence of un-operated cataracts. In developing countries such as Nigeria, the onset of cataract blindness can often be much earlier [9]. Risk factors that predispose an individual to cataract are: age over 40 years, poorly controlled diabetes, smoking, use of steroids, trauma to the eye, positive family history, exposure to ultra-violet light from the sun and x-rays [7].

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 [10]. Accurate predictive models can inform patients and physicians about the future course of an illness or the risk of developing illness thereby guiding decisions on screening and/or treatment [11]. Data Mining or the efficient discovery of valuable on-obvious information from a large collection of data [12] and has a goal to discover knowledge out of data analyzed. Knowledge discovery in database is a precise process consisting of number of distinct steps including data mining [13].

Data mining has a great potential to enable healthcare systems to make data more efficient and effective thereby reducing the likely costs associated with making decisions. Also, data mining techniques are very useful in healthcare domain because they provide better medical services to the patients and helps healthcare organizations in various medical management decisions [14]. With the help of classification approach, a risk factor can be associated to patients by analyzing their patterns of diseases. Machine learning algorithms provide means of obtaining objective unseen patterns from evidence-based information especially in the public health care sector. With aging populations more disposed to sight loss and risks such as diabetes increasing, levels of avoidable blindness in the region are likely to rise. Strategic approaches addressing the eye health workforce crisis are now urgent and essential. In Africa, there are major data gaps in eye health on prevalence of conditions, personnel, access to services, availability of rehabilitation and assistive devices. In Nigeria today, the risk of cataract is difficult to pre-determine before the onset of the disease, and most times individuals are already showing signs of blindness by the time they reach the doctors for complaints. Developed nations all over have systems in place that aid the early detection of related eye diseases but developing nations like Nigeria lack facilities that possess such capability. There is a need for a model that can be used to determine the risk of cataract in Nigerians so as to help reduce the number of blindness cases in Nigeria, hence this study.

2. RELATED WORKS

Bowd and Goldbaum [15] worked on the application of machine learning classifiers for the detection of glaucoma. The study provided some background about the classification task in glaucoma and the structure and evaluation of Machine learning classifiers (MLCs), and it reviews MLC techniques as they have been applied to visual function and optical imaging in glaucoma research. Imberman [16] 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 that 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 diseases. The results showed that the model developed had specificity of 37% and sensitivity of 98%.

Kabari and Nwachukwu [17] 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. From the 400 data samples, 320 samples (80%) were randomly chosen and used as training patterns, while 80 instances (20%) of the same data set were used for testing. The data set consists of evenly distributed men and women. Samples also considered age randomly 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 implies that combination of Neural Networks and decision Tree Technique is an effective and efficient method for implementing diagnostic problem. Umesh *et al.* [18] 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 and 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.

3. METHODS

The methodological approach of this study composes of a number of methods namely: the identification of the required variables for the risk of cataracts, the collection of historical datasets about cataracts risk cases of patients, formulation of the predictive models using the supervised machine learning algorithms proposed, the simulation of the predictive models using the WEKA simulation environment and the performance evaluation metrics applied during model validation of the predictive models. The supervised machine learning algorithms chosen for this study are C4.5 decision trees, Naïve Bayes and Multi-layer Perceptron.

3.1 Data Collection

For the purpose of this study, data was collected from 30 patients located in the south-western part of Nigeria using structured questionnaires that consisted of two (2) main sections, namely: Demographic Factor and Clinical Factor with the former having gender, age, marital status, ethnicity, occupation, religion and academic qualification as demographic information the latter uses axial length of the eye, family history of cataract, height (meters). Weight (Kg), and other clinical variables necessary for identifying the risk of cataract. The information collected consisted of the risk factors associated with the cataracts for each patient as proposed by the ophthalmologist. A description of the attributes contained in the dataset is presented in Table 1.

Table 1: Identified Variables for the Risk of Cataracts

Categories	Risk Factors	Labels
Demographic	**Age (years)**	< 10, 10 – 18, 19 – 35, > 35
	Education	Primary, Secondary, University, Polytechnic
	Occupation	Trader, Teacher, Student, Artisan, Military, Self, Civil-servant
	Marital Status	Married, Single
	Ethnicity	Yoruba, Hausa, Ibo
	Religion	Christianity, Islam
Clinical	**Weight (Kg)**	Numeric
	Height (metres)	Numeric
	BMI Class	Underweight, Normal, Obese, Overweight
	Use Lenses	Yes, No
	Family History	None, First, second
	Smoking	Yes, No
	Smoke Frequency	No, pack/day, pack/week, pack/month, pack/year
	High Cholesterol	Yes, No
	Diabetes	Yes, No
	Hypertension	Yes, No
	Hypertensive	Yes, No
	Corticosteroid Medications	Yes, No
	Past Eye Surgery	Yes, No
	Hormone Replacement	Yes, No
	Have Myopia	Yes, No
	Myopia Intensity	None, Low, Moderate, High
	Alcohol	Yes, No
	Alcohol Frequency	None, Monthly, Daily, Weekly
Target Class	**Risk of Cataracts**	No, Low, Moderate, High

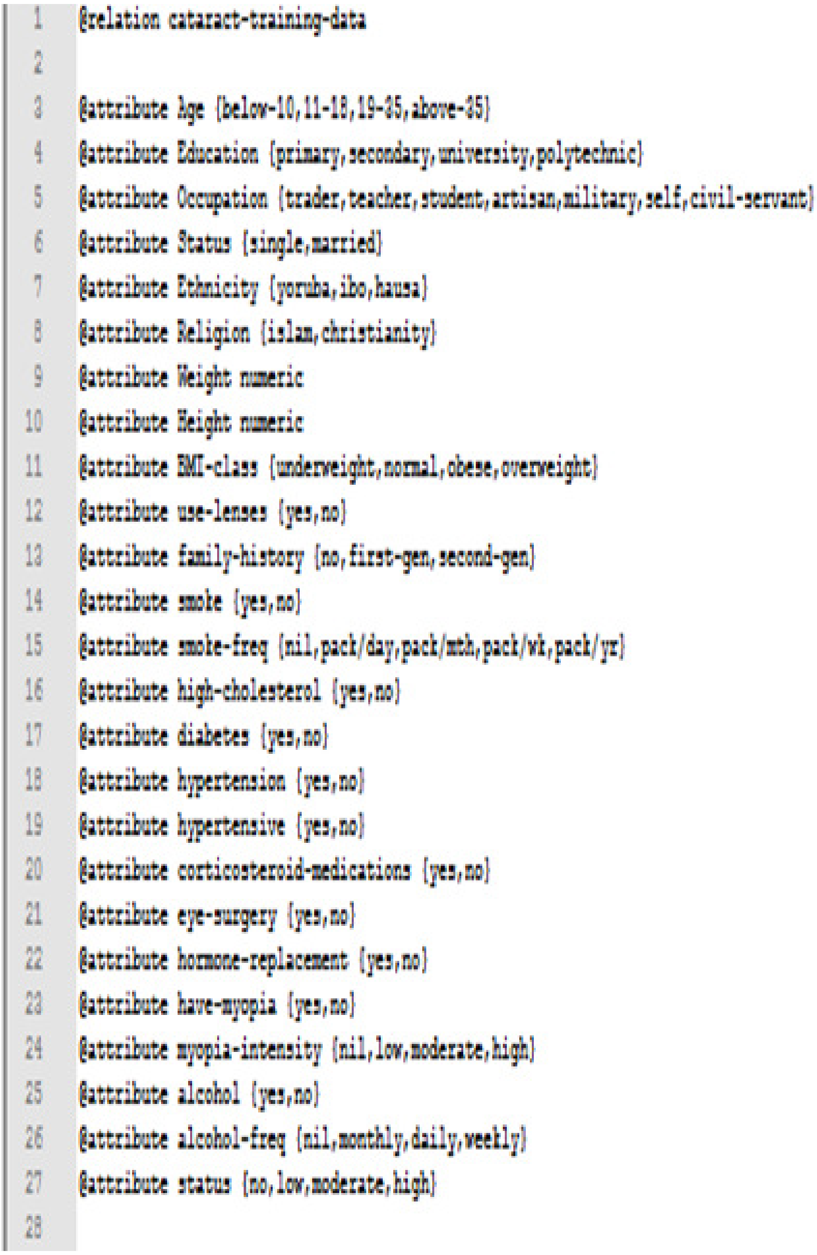
3.2 Data-Preprocessing

Following the collection of data from the 31 patients alongside the attributes (24 risk factors) alongside the risk of cataracts, the data collected was checked for the presence of error in data entry including misspellings and missing data. Following this process, there was no error in misspellings but there were missing data in the cells describing some records. The data was transformed into the attribute file format (.arff) for the purpose of the development of the predictive model for the risk of cataracts using the simulation environment. Figure 1 shows a screenshot of the format of the .arff used for model development in the Waikato Environment for Knowledge Analysis (WEKA) – a light-weight java application composed of a suite of supervised and unsupervised machine learning tools. The arff file is composed of three parts:

a. The relation name section, which contains the tag @relation *cataract-training-data* used to identify the name of the relation (or file) that contains the data needed for simulation. This section is located at the first line of the file and the tag 'name' following @relation must always be the same as the file name else the file loader of the simulation environment will cease to open the file. This section is followed by the attribute names section;

b. The attribute names section, which contains the tag @attribute *attribute_name label* used to identify the attributes that describe the dataset stored in the .arff file needed for simulation. Each attribute name alongside its labels is stated following the @relation tag on each line. The label can be a set of values inserted between brackets or a descriptor (e.g. date, numeric etc.). The last attribute is identified as the target class (risk of cataracts) while the previous attributes are the risk factors for the risk of infertility. Also, this section is followed by the data section.

c. This contains the tag @data followed in the next line by the values of the attributes for each record of the risk of infertility separated by a comma. Each value was listed on a row for each record in the same order as the attributes were listed in the attribute names section. The values inserted into each record must be the same values defined in each respective attribute; if there is an error in spelling or a label is not defined and inserted then the file loader of the simulation environment will fail to load the file.

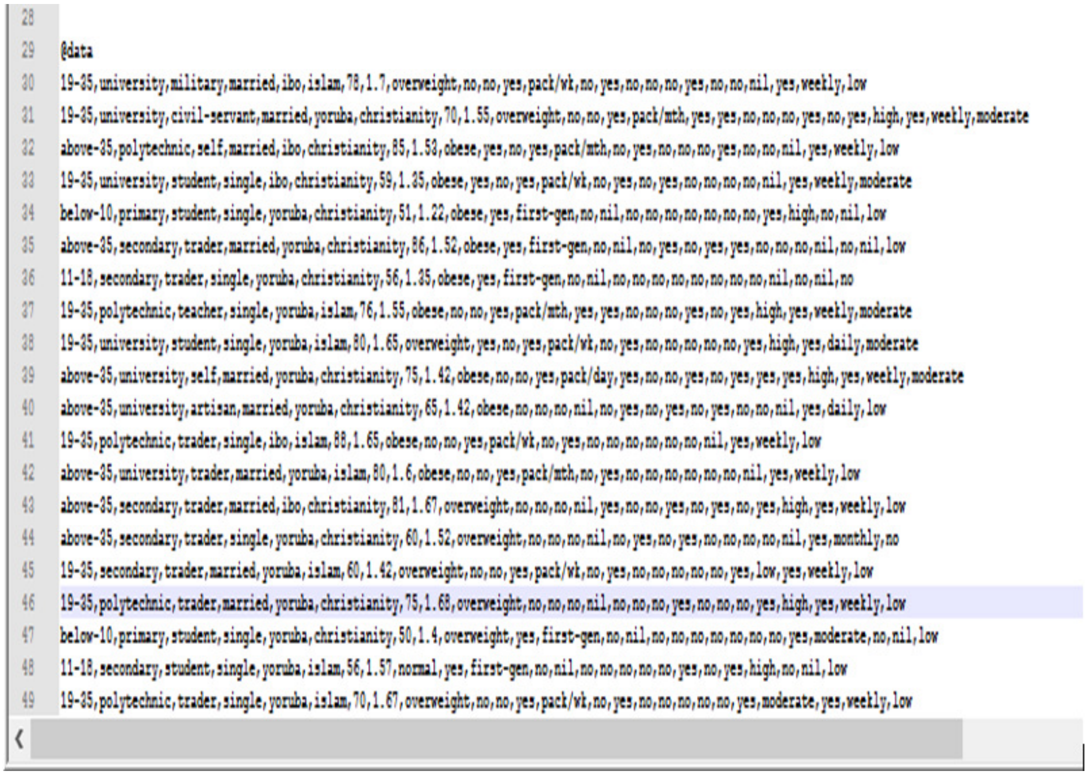


```
@relation cataract-training-data

@attribute Age {below-10,11-18,19-35,above-35}
@attribute Education {primary,secondary,university,polytechnic}
@attribute Occupation {trader,teacher,student,artisan,military,self,civil-servant}
@attribute Status {single,married}
@attribute Ethnicity {yoruba,ibo,hausa}
@attribute Religion {islam,christianity}
@attribute Weight numeric
@attribute Height numeric
@attribute BMI-class {underweight,normal,obese,overweight}
@attribute use-lenses {yes,no}
@attribute family-history {no,first-gen,second-gen}
@attribute smoke {yes,no}
@attribute smoke-freq {nil,pack/day,pack/mth,pack/wk,pack/yr}
@attribute high-cholesterol {yes,no}
@attribute diabetes {yes,no}
@attribute hypertension {yes,no}
@attribute hypertensive {yes,no}
@attribute corticosteroid-medications {yes,no}
@attribute eye-surgery {yes,no}
@attribute hormone-replacement {yes,no}
@attribute have-myopia {yes,no}
@attribute myopia-intensity {nil,low,moderate,high}
@attribute alcohol {yes,no}
@attribute alcohol-freq {nil,monthly,daily,weekly}
@attribute status {no,low,moderate,high}

```

Figure 1a: arff file containing identified attributes



```

@data
19-35,university,military,married,ibo,islam,78,1.7,overweight,no,no,yes,pack/wk,no,yes,no,no,no,yes,no,no,nil,yes,weekly,low
19-35,university,civil-servant,married,yoruba,christianity,70,1.55,overweight,no,no,yes,pack/mth,yes,yes,no,no,no,yes,no,yes,high,yes,weekly,moderate
above-35,polytechnic,self,married,ibo,christianity,85,1.53,obese,yes,no,yes,pack/mth,no,yes,no,no,no,yes,no,no,nil,yes,weekly,low
19-35,university,student,single,ibo,christianity,59,1.35,obese,yes,no,yes,pack/wk,no,yes,no,yes,no,no,no,no,nil,yes,weekly,moderate
below-10,primary,student,single,yoruba,christianity,51,1.22,obese,yes,first-gen,no,nil,no,no,no,no,no,no,no,yes,high,no,nil,low
above-35,secondary,trader,married,yoruba,christianity,86,1.52,obese,yes,first-gen,no,nil,no,yes,no,yes,yes,no,no,no,nil,no,nil,low
11-18,secondary,trader,single,yoruba,christianity,56,1.35,obese,yes,first-gen,no,nil,no,no,no,no,no,no,no,no,nil,no,nil,no
19-35,polytechnic,teacher,single,yoruba,islam,76,1.55,obese,no,no,yes,pack/mth,yes,yes,no,no,no,yes,no,yes,high,yes,weekly,moderate
19-35,university,student,single,yoruba,islam,80,1.65,overweight,yes,no,yes,pack/wk,no,yes,no,no,no,no,no,yes,high,yes,daily,moderate
above-35,university,self,married,yoruba,christianity,75,1.42,obese,no,no,yes,pack/day,yes,no,no,yes,no,yes,yes,yes,high,yes,weekly,moderate
above-35,university,artisan,married,yoruba,christianity,65,1.42,obese,no,no,no,nil,no,yes,no,yes,no,yes,no,no,nil,yes,daily,low
19-35,polytechnic,trader,single,ibo,islam,88,1.65,obese,no,no,yes,pack/wk,no,yes,no,no,no,no,no,no,nil,yes,weekly,low
above-35,university,trader,married,yoruba,islam,80,1.6,obese,no,no,yes,pack/mth,no,yes,no,no,no,no,no,no,nil,yes,weekly,low
above-35,secondary,trader,married,ibo,christianity,81,1.67,overweight,no,no,no,nil,yes,no,no,yes,no,yes,no,yes,high,yes,weekly,low
above-35,secondary,trader,single,yoruba,christianity,60,1.52,overweight,no,no,no,nil,no,yes,no,yes,no,no,no,no,nil,yes,monthly,no
19-35,secondary,trader,married,yoruba,islam,60,1.42,overweight,no,no,yes,pack/wk,no,yes,no,no,no,no,no,yes,low,yes,weekly,low
19-35,polytechnic,trader,married,yoruba,christianity,75,1.68,overweight,no,no,no,nil,no,no,no,yes,no,no,no,yes,high,yes,weekly,low
below-10,primary,student,single,yoruba,christianity,50,1.4,overweight,yes,first-gen,no,nil,no,no,no,no,no,no,no,yes,moderate,no,nil,low
11-18,secondary,student,single,yoruba,islam,56,1.57,normal,yes,first-gen,no,nil,no,no,no,no,no,yes,no,yes,high,no,nil,low
19-35,polytechnic,trader,single,yoruba,islam,70,1.67,overweight,no,no,yes,pack/wk,no,yes,no,no,no,no,no,yes,moderate,yes,weekly,low
```

Figure 1b: arff file containing identified attributes (contd)

The dataset collected for the purpose of the development of the predictive model for the risk of cataracts was stored in .arff in the name *cataract-training-data.arff* while the number of attributes listed in the attribute section were 25 including the target attribute. Following this, the values of the risk factors for the record of the 31 patients considered for this study were provided.

3.3 Model Formulation

Systems that construct classifiers are one of the commonly used tools in data mining. Such systems take input of collected cases with each belonging to one of a small number of classes and described by its values for a fixed set of attributes, and output a classifier that can accurately predict the class to which a new case belongs. Supervised machine learning algorithms makes it possible to assign a set of records (cataracts risk indicators) to a target classes – the risk of cataracts. Equation 1 shows the mapping function that describes the relationship between the risk factors and the target class – risk of cataracts.

$$\varphi: X \rightarrow Y \qquad (1)$$
$$defined\ as: \varphi(X) = Y$$

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 cataracts – low, moderate and high risk of cataracts 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 cataracts risk maps the risk factors of each individual to their respective risk of cataracts according to equation 2.

$$\varphi(X) = \begin{cases} No\ Risk \\ Low\ Risk \\ Moderate\ Risk \end{cases} \qquad (2)$$

Supervised machine learning algorithms are Black-boxed models, thus it is not possible to give an exact description of the mathematical relationship existing among the independent variables (input variables) with respect to the target variable (output variable – risk of cataracts). Cost functions are used by supervised machine learning algorithms to estimate the error in prediction during the training of data for model development. Although, the decision trees algorithm is a white-boxed model owing to its ability of being interpreted as a tree-structure.

3.3.1 Naïve Bayes' Classifier

Naive Bayes' Classifier is a probabilistic model based on Bayes' theorem. It is defined as a statistical classifier. It is one of the frequently used methods for supervised learning. It provides an efficient way of handling any number of attributes or classes that are purely based on probabilistic theory. Bayesian classification provides practical learning algorithms and prior knowledge on observed data. Let X_{ij} be a dataset sample containing records (or instances) of *i* number of risks factors (attributes/features) alongside their respective risk of cataracts, *C* (target class) collected for *j* number of records/patients and $H_k = \{H_1 = No, H_2 = Low, H_3 = Moderate\}$ be a hypothesis that X_{ij} belongs to class C. For the classification of the risk of infertility given the values of the risk factor of the jth record, Naïve Bayes' classification required the determination of the following:

a. $P(H_k|X_{ij})$ – Posteriori probability: is the probability that the hypothesis, H_k holds given the observed data sample X_{ij} for $1 \leq k \leq 3$.
b. $P(H_k)$ - Prior probability: is the initial probability of the target class $1 \leq k \leq 3$;
c. $P(X_{ij})$ is the probability that the sample data is observed for each risk factor (or attribute), *i*; and
d. $P(|X_{ij}|H_k)$ is the probability of observing the sample's attribute, X_i given that the hypothesis holds in the training data X_{ij}.

Therefore, the posteriori probability of an hypothesis H_k is defined according to Bayes' theorem as follows:

$$P(H_k|X_{ij}) = \frac{\prod_{i=1}^{n} P(X_{ij}|H_k)P(X_i)}{P(H_k)} \quad for\ k = 1,2,3 \qquad (3)$$

Hence, the risk of cataracts for a record is thus:

$$Risk = MAX[P(no|X_k).P(low|X_k).P(moderate|X_k)] \qquad (4)$$

3.3.2 Decision Trees Algorithm

The theory of a decision tree has a root node which is the starting point of the tree and branches which connect nodes showing the flow from question to answer. Nodes that have child nodes are called interior nodes. Leaf or terminal nodes are nodes that do not have child nodes and represent a possible value of target variable given the variables represented by the path from the root. The rules are inducted by definition from each respective node to branch to leaf**.** Given a set X_{ij} of *j* number of cases, the decision trees algorithm grows an initial tree using the divide-and-conquer algorithm as follows:

a. If all the cases in X_{ij} belonging to the same class and X_{ij} is small, the tree is a leaf labeled with the most frequent class in X_{ij}.
b. Otherwise, choose a test based on a single attribute X_i with two or more outcomes. Make this test the root of the tree with one branch for each outcome of the test, partition X_{ij} into corresponding subsets according to the outcome for each case, and apply the same procedure recursively to each subset.

ID3 (Iterative Dichotomiser 3) decision trees algorithm is a classification tree used in the concept of information entropy which provided a method for measuring the number of bits each attribute can provide. Hence, the attribute that yields the most information gain becomes the most important attribute and should thus go to the top of the tree. The C4.5 decision trees algorithm builds decision trees from a set of training dataset, X_{ij} the same way as ID3, using the information entropy. For this study, the C4.5 decision trees algorithm was used for the formulation of the predictive model for the risk of cataracts due to its advantages over the ID3 decision trees algorithm and due to its ability to handle continuous and discrete attributes, missing values, attributes with differing costs and prune trees after creation.

The two criteria used by the C4.5 decision trees in developing its decision trees are presented in equations (4) and (5) defined as the information gain and the split criteria respectively. Equation (4) is used in determining which attribute is used to split the dataset at every iteration while equation (5) is used to determine which of the selected attribute split is most effective in splitting the dataset after attribute selection by equation (4).

$$IG(X_i) = H(X_i) - \sum_{t \epsilon T} \frac{|t|}{|X_{i\square}|} \cdot H(X_i) \quad (4)$$

Where:

$$H(X_i) = -\sum_{t \epsilon T} \frac{|t, X_i|}{|X_{ij}|} \cdot \log_2 \frac{|t, X_i|}{|X_{ij}|}$$

$$Split(T) = -\sum_{t \epsilon T} \frac{|t|}{|X_{ij}|} \cdot \log_2 \frac{|t|}{|X_{ij}|} \quad (5)$$

T is the set of values for a given attribute X_i.

3.3.3 Multi-Layer Perceptron

An artificial neural network (ANN) is an interconnected group of nodes akin to the vast network of neurons in a human brain. Multi-layer perceptron are ANNs which are generally presented as systems of interconnected neurons (containing activation functions) which send messages to each other such that each connection has numeric weights that can be tuned based on experience, making neural nets adaptive to inputs and capable of learning using the back-propagation algorithm. For this study, the input variables (risk factors of cataract risk) were fed to the MLP as inputs to which initially random values within the interval [0, 1] were assigned. Each weight was assigned to their respective inputs as shown in equation (6) and propagated through the activation function of each neuron in the hidden layers of the MLP architecture shown in equation (7).

$$\sum_{k=1}^{i} w_k x_k = w_1 x_1 + w_2 x_2 + \cdots + w_i x_i = < w.x > \quad (6)$$

Using the back-propagation algorithm, the MLP compares the output calculated with the actual in order to compute an error-function. Gradient descent was then used to feed the error back to the system from output nodes through the nodes in the hidden layers to the nodes at the input layer while adjusting the weights as a function of the error determined at each node. The process was repeated for a number of training cycles for which the MLP network converged to a state where the error determined is small enough, then the MLP network was able to learn the target function.

The back-propagation learning algorithm can be divided into two phases: propagation and weight update.

a. Phase 1 – Propagation: each propagation involves the following steps:

   i. Forward propagation of training pattern's input through each node *j* in the neural network in order to generate the propagation's output activations;

$$output\ O_j = \varphi(\sum_{k=1}^{i} w_{kj} x_k + b_k) = \varphi(z) = \frac{1}{1 + e^{-z}} \quad (7)$$

   ii. Backward propagation of the propagation's output activations through the neural network using the training pattern target in order to generate deltas δ_j of all output and hidden neurons.

$$\delta_j = \frac{\partial E}{\partial O_j} \frac{\partial O_j}{\partial net_j} = \begin{cases} (O_j - p_j)\varphi(net_j)\left(1 - \varphi(net_j)\right) j\ is\ output\ neuron, \\ (\sum_{l \epsilon L} \delta_j w_{jl})\varphi(net_j)\left(1 - \varphi(net_j)\right) j\ is\ inner\ neuron \end{cases} \quad (8)$$

b. Phase 2 – Weight update: for each weight-synapse, hence the following:

   i. Multiply its output delta and input activation to get the gradient of the weight

$$\frac{\partial E}{\partial w_{ij}} = \delta_j x_i \quad (9)$$

   ii. Subtract a ratio (percentage α) of the gradient from the weight.

$$\Delta w_{ij} = -\alpha \frac{\partial E}{\partial w_{ij}} \quad (10)$$

3.4 Performance Evaluation

In order to evaluate the performance of the supervised machine learning algorithms used for the classification of the risk of cataracts, there was the need to plot the results of the classification on a confusion matrix (Figure 2). A confusion matrix is a square which shows the actual classification along the vertical and the predicted along the horizontal. Correct classifications were plotted along the diagonal from the north-west position for the low cases predicted as No (A), low (E) and moderate (I) on the south-east corner (also called true positives and negatives). The incorrect classifications were plotted in the remaining cells of the confusion matrix (also called false positives). . These results are presented on confusion matrix – for this study the confusion matrix is a 3 x 3 matrix table owing to the three (3) labels of the output class.

NO	LOW	MODERATE	
A	B	C	NO
D	E	F	LOW
G	H	I	MODERATE

Figure 2: Diagram of a Confusion Matrix

Also, the actual no cases are A+B+C, actual low cases are D+E+F and the actual moderate cases are G+H+I while the predicted no are A+D+G, predicted low are B+E+H and predicted moderate are C+F+I. The developed model was validated using a number of performance metrics based on the values of *A – I* in the confusion matrix for each predictive model. They are presented as follows.

a. Accuracy: the total number of correct classification

$$Accuracy = \frac{A + E + I}{total\_cases} \qquad (11)$$

b. True positive rate (recall/sensitivity): the proportion of actual cases correctly classified

$$TP_{no} = \frac{A}{A + B + C} \qquad (12)$$

$$TP_{low} = \frac{E}{D + E + F} \qquad (13)$$

$$TP_{moderate} = \frac{I}{G + H + I} \qquad (14)$$

c. False positive (false alarm/1-specificity): the proportion of negative cases incorrectly classified as positive

$$FP_{no} = \frac{D + G}{actual_{low} + actual_{moderate}} \qquad (15)$$

$$FP_{low} = \frac{B + H}{actual_{no} + actual_{moderate}} \qquad (16)$$

$$FP_{moderate} = \frac{C + F}{actual_{no} + actual_{low}} \qquad (17)$$

d. Precision: the proportion of predictions that are correct

$$Precision_{no} = \frac{A}{A + D + G} \qquad (18)$$

$$P\square ecision_{low} = \frac{F}{B + E + H} \qquad (19)$$

$$Precision_{moderate} = \frac{K}{C + F + I} \qquad (20)$$

4. RESULTS

This section presents the results of the methods that were applied for the development of the predictive model for the risk of cataracts. The results presented were that of the data collection, model formulation and simulation results using the WEKA software following the results of the model validation of the predictive model for cataracts.

4.1 Data Description

For this study, data was collected from 30 patients using the questionnaires constructed for this study among which; the risk of cataracts was identified.

Table 2: Distribution of cataracts risk among historical dataset

Cataract risk	Frequency	Percentage (%)
No	7	22.58
Low	18	58.07
Moderate	6	19.35
Total	31	100.00

Table 2 above gives a description of the number of patients with their respective risk of cataracts from the 31 patient records selected for model formulation and validation which were stored in the cataract-training-data file. The table shows that out of the 31 patients considered; 22.6% of the respondents had no risk of cataracts, 58.1% of the respondents had low risk of cataracts while 19.3% of respondents had moderate risk of cataracts. It was observed that the highest case presented was for respondents with low risk of cataracts while the least case was presented for respondents with moderate risk of cataracts.

Table 3: Description of Demographic Data of respondents

Variable Name	Labels	Frequency	Percentage (%)
Age (years)	Below	2	6.5
	11-18	3	9.7
	19-35	16	51.6
	Above 35	10	32.3
Education	Primary	2	6.5
	Secondary	9	29.0
	Polytechnic	7	22.6
	University	13	41.9
Marital Status	Single	15	48.4
	Married	16	51.6
Ethnicity	Yoruba	22	71.0
	Ibo	9	29.0
Occupation	Civil servant	1	3.2
	Trader	15	48.4
	Teacher	3	9.7
	Student	8	25.8
	Artisan	1	3.2
	Military	1	3.2
	Self	2	6.4
Religion	Christian	20	64.5
	Islam	11	53.5
BMI Class	Normal	3	9.7
	Overweight	12	38.7
	Obese	16	51.6

Table 3 shows a description of the demographic data collected from all 31 respondents selected for the study. It also shows the distribution of the values of each demographic variable considered. From the data presented, a number of results were observed as presented in the following paragraphs. Less than 20% of the respondents were below 19 years of age with 51.6% of the respondents within the age group of 19 – 35 years and the remaining 32.3% within the age group of 35 years and above. About 41.9% of the respondents had attended the university while 51.6% of the respondents had either attended secondary schools or polytechnic while about 51.6% and 48.4% of the respondents were married and single respectively.

About 48.4% of the respondents selected for this study were traders while the remaining 51.6% consisted of civil servants, teachers, students, self-employed, artisans and military. There were more Christians than Muslims among the respondents selected for this study while about 50% of the respondents selected were obese.

Table 4: Description of Risk Factor Data of respondents

Risk Factor Information	Labels	Frequency	Percentage (%)
Use Lenses	Yes	8	25.8
	No	23	74.2
Family History	None	25	80.6
	First	6	19.4
	Second	0	0.0
Smoking	Yes	15	48.4
	No	16	51.6
Smoke Frequency	No	16	51.6
	Pack/day	1	3.2
	Pack/week	8	25.8
	Pack/month	6	19.4
	Pack/year	0	0.0
High Cholesterol	Yes	8	25.8
	No	23	74.2
Diabetes	Yes	16	51.6
	No	15	48.4
Hypertension	Yes	1	3.2
	No	30	96.8
Hypertensive	Yes	9	29.0
	No	22	71.0
Corticosteroid Medications	Yes	3	9.7
	No	28	90.3
Past Eye Surgery	Yes	12	38.7
	No	19	61.3
Hormone Replacement	Yes	1	3.2
	No	30	96.8
Have Myopia	Yes	17	54.8
	No	14	45.2
Myopia Intensity	None	14	45.2
	Low	1	3.2
	Moderate	6	19.4
	High	10	32.3
Alcohol	Yes	21	67.7
	No	10	32.3
Alcohol Frequency	None	9	29.0
	Monthly	5	16.1
	Daily	2	6.5
	Weekly	15	48.4

Table 4 on tne previous page shows the description of the risk factor information for all 31 patients alongside their frequency distribution. The table displays different responses of the respondents selected for this study regarding the risk factors of cataracts. The description of the responses regarding the risk factors were presented using frequency distribution tables. The results of distribution show responses of the 31 respondents regarding the risk factors of cataracts and a number of deductions were made. The results showed that (74.2%) of majority of the respondents had used lenses; (80.6%) had no family history of cataracts; (51.6%) had not been smoking; (74.2%) had no high cholesterol level; (51.6%) had diabetes; (96.8%) had hypertension but (71%) not hypertensive; (90.3%) did not take corticosteroid medications; (61.3%) had no eye surgery in the past; (96.8%) had no hormone replacement have myopia; (54.8%) out of which majority were high (32.3%) and (67.7%) were alcoholic among whom (48.4%) indulged weekly.

4.2 Simulation Results

Three different supervised machine learning algorithms were used to formulate the predictive model for the risk of cataracts, namely: Naïve Bayes', Decision Trees and the Multi-layer Perceptron classifiers. They were used to train the development of the prediction model using the dataset containing 31 patients' risk factor records. The simulation of the prediction models was done using the Waikato Environment for Knowledge Analysis (WEKA). The C4.5 Decision Trees Algorithm was implemented using the J48 available in the trees class, the naïve Bayes' algorithm was implemented using the naïve Bayes' classifier available in the Bayes class and the multi-layer perceptron was implemented using the multilayer perceptron class all available on the WEKA environment of classification tools. The models were trained using the 10-fold cross validation method that split the dataset into 10 subsets of data – while 9 parts are used for training, the remaining one is used for testing; this process is repeated until the remaining 9 parts take their turn for testing the model.

4.2.1 Results of the C4.5 decision trees classifier

Using the C4.5 decision trees classifier to train the predictive model developed using the training data via the 10-fold cross validation method. Figure 3 shows the graphical plot of the predictions made by the C4.5 decision trees classifier algorithm on the dataset, each class of cataracts is represented using a specific colour and each correct classification is represented with a star while each misclassification is represented as a square. The results presented in figure 3 were used to evaluate the performance of the C4.5 decision trees classifier algorithm and thus, the confusion matrix determined as shown in figure 4. From the confusion matrix shown in figure 4, the following sections present the results of the model's performance. Out of the 7 no actual cases, 5 were correctly classified as no while 2 were misclassified as low risk, out of the 18 actual low cases, there were 17 correct classifications with 1 misclassified as no risk and out of the 6 moderate risk cases, there were 5 correct classifications with 1 misclassified as low risk. Therefore, there were 27 correct classifications out of the 31 records considered for the model development owing for an accuracy of 87.1%.

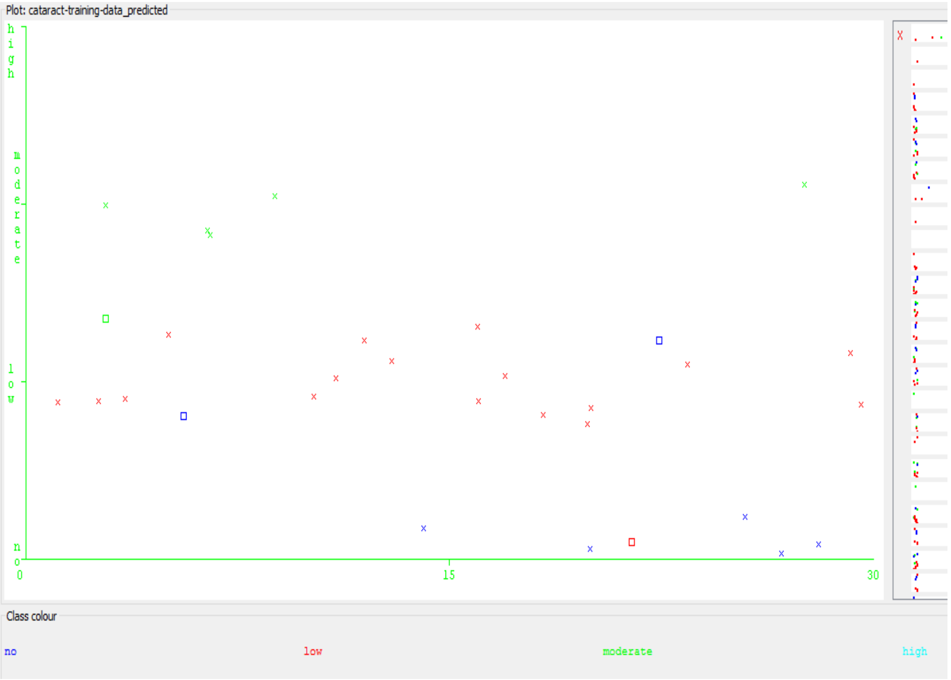


Figure 3: Screenshot of C4.5 decision trees Classification Results

NO	LOW	MODERATE	
5	2	0	NO
1	17	0	LOW
0	1	5	MODERATE

Figure 4: Confusion matrix of performance evaluation using C4.5 decision trees

The decision tree that was plotted from the simulation of the predictive model using the C4.5 decision trees is presented in figure 5. The main risk factors considered with the highest gain ratio by C4.5 were smoking, myopia intensity, alcohol frequency and use of lenses. Using the decision tree in Figure 5, the following rules were deduced and can be used to predict the risk of cataracts based on the values of the three identified risk factors. There are 9 rules and these are presented as follows:

i. If (smoking = YES) AND (myopia intensity = NIL) Then (cataract risk = LOW);
ii. If (smoking = YES) AND (myopia intensity = LOW) Then (cataract risk = LOW);
iii. If (smoking = YES) AND (myopia intensity = MODERATE) Then (cataract risk = LOW);
iv. If (smoking = YES) AND (myopia intensity = HIGH) Then (cataract risk = MODERATE);
v. If (smoking = NO) AND (alcohol frequency = NIL) AND (use lenses = YES) then (cataract risk = LOW);
vi. If (smoking = NO) AND (alcohol frequency = NIL) AND (use lenses = NO) then (cataract risk = NO);
vii. If (smoking = NO) AND (alcohol frequency = MONTHLY) then (cataract risk = NO);
viii. If (smoking = NO) AND (alcohol frequency = DAILY) then (cataract risk = LOW); and
ix. If (smoking = NO) AND (alcohol frequency = DAILY) then (cataract risk = WEEKLY).

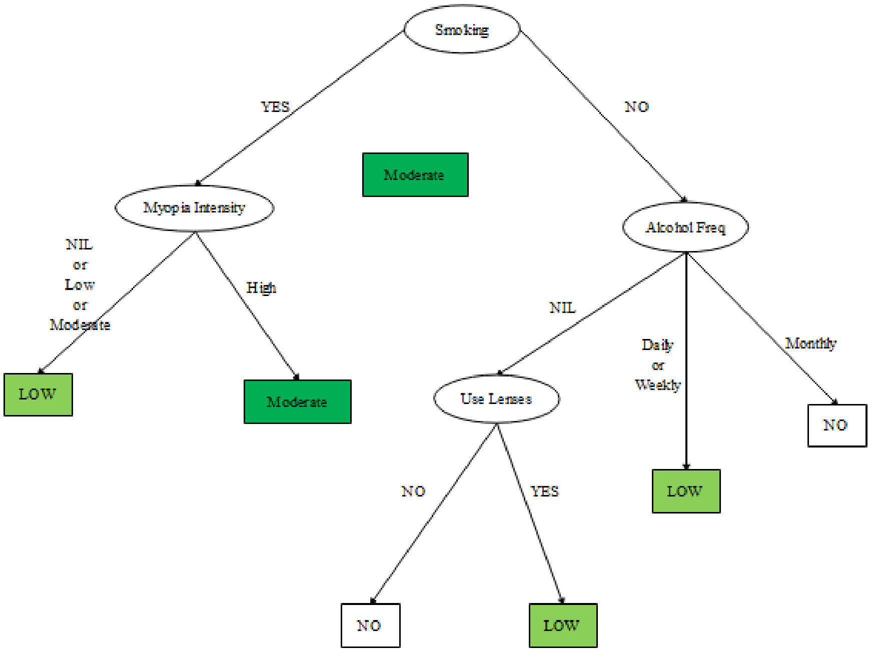


Figure 5: Graphical plot of the C4.5 decision tree for cataracts risk

4.2.2 Results of the naïve Bayes' classifier

Using the Naïve Bayes' classifier to train the predictive model developed using the training data via the 10-fold cross validation method. Figure 6 shows the graphical plot of the predictions made by the Naive Bayes classifier algorithm on the dataset, each class of cataracts is represented using a specific colour and each correct classification is represented with a star while each misclassification is represented as a square. The results presented in figure 6 were used to evaluate the performance of the Naive Bayes classifier algorithm and thus, the confusion matrix determined as shown in figure 7. From the confusion matrix shown in figure 7, the following sections present the results of the model's performance. Out of the 8 actual no cases, all were correctly classified, out of the 7 actual low cases, there were 4 correct classifications with 2 misclassified as no risk and 1 misclassified as high risk; out of the 7 moderate risk cases, there were 5 correct classifications with 1 misclassified as no risk and 1 misclassified as low risk while out of the 8 high cases, there were 6 correct classifications with 2 misclassified as low risk. Therefore, there were 23 correct classifications out of the 30 records considered for the model development owing for an accuracy of 76.67%.

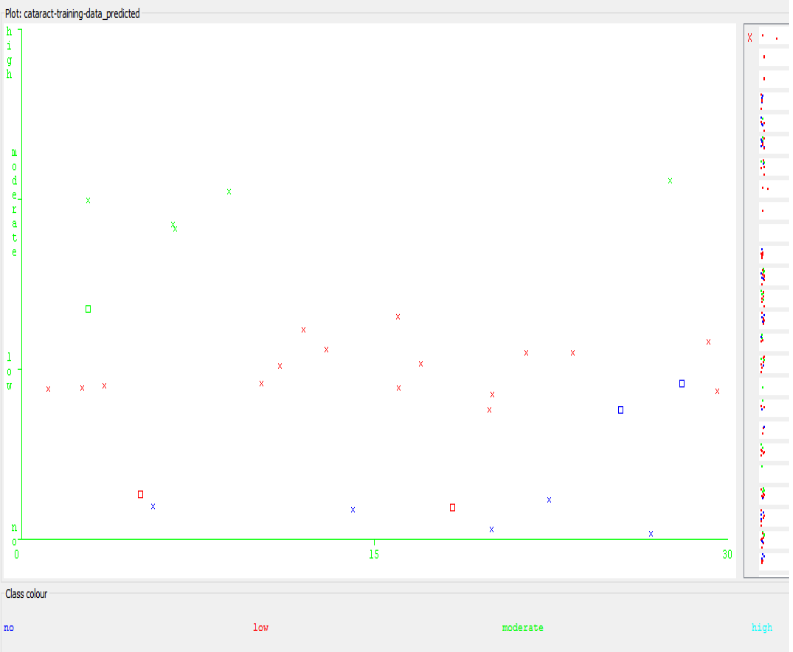


Figure 6: Screenshot of Naïve Bayes' Classification Results

NO	LOW	MODERATE	
5	2	0	NO
2	16	0	LOW
0	1	5	MODERATE

Figure 7: Confusion matrix of performance evaluation using naïve Bayes

4.2.2 Results of the Multi-layer perceptron classifier

Using the Multi-layer perceptron classifier to train the predictive model developed using the training data via the 10-fold cross validation method. Figure 8 shows the graphical plot of the predictions made by the Multi-layer perceptron classifier algorithm on the dataset, each class of cataracts is represented using a specific colour and each correct classification is represented with a star while each misclassification is represented as a square. The results presented in figure 8 were used to evaluate the performance of the Multi-layer Perceptron classifier algorithm and thus, the confusion matrix was determined as shown in figure 9.

From the confusion matrix shown in figure 9, the following sections present the results of the model's performance. Out of the 7 actual no cases, all were correctly classified, out of the 18 actual low cases, all were correctly classified and out of the 6 moderate risk cases, all were correctly classified. Therefore, all 31 records considered for the model development were all correctly classified indicating an accuracy of 100%.

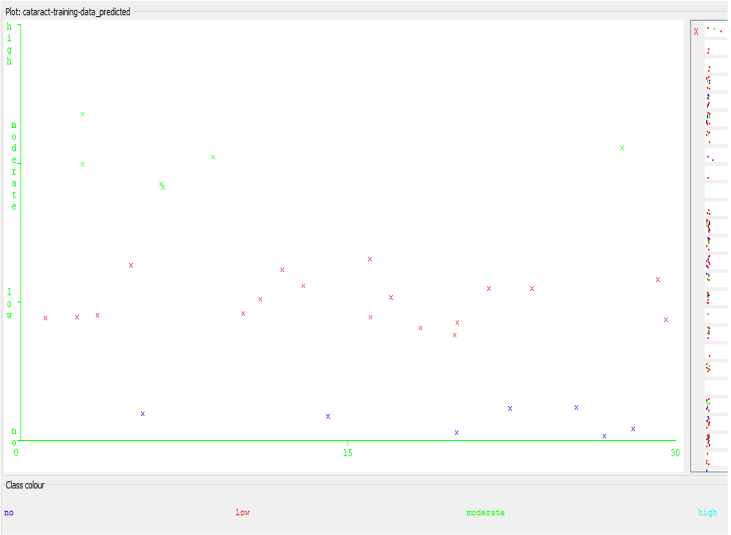


Figure 8: Screenshot of Multi-layer perceptron Classification Results

NO	LOW	MODERATE	
7	0	0	NO
0	18	0	LOW
0	0	6	MODERATE

Figure 9: Confusion matrix of performance evaluation using Multi-layer perceptron

4.3 Discussions

The result of the performance evaluation of the machine learning algorithms is presented in Table 5, which presents the average values of each performance evaluation metrics considered for this study.

Table 5: Summary of Validation Results for C4.5, naïve Bayes' and MLP classifiers

Machine Learning Algorithm Used	PERFORMANCE EVALUATION METRICS				
	Correct Classification (out of 45)	Accuracy (%)	TP rate (recall or sensitivity)	FP rate (false positive)	Precision
C4.5 Decision Trees Algorithm	27	87.1	0.830	0.091	0.894
Naïve Bayes' Classifier	26	83.9	0.812	0.105	0.852
Multi-Layer Perceptron Algorithm	**31**	**100.0**	**1.000**	**0.000**	**1.000**

For the C4.5 Decision Trees algorithm based on the results presented in the confusion matrix presented in figure 4. The result showed that the TP rate which gave a description of the proportion of actual cases that was correctly predicted was 0.830, which implied that 83% of the actual cases were correctly predicted; the FP rate which gave a description of the proportion of actual cases misclassified was 0.091. This implied that 9% of actual cases were misclassified while the precision which gave a description of the proportion of predictions that were correctly classified was 0.894, which implied that 89% of the predictions made by the classifier were correct.

For the Naive Bayes classifier algorithm based on the results presented in the confusion matrix presented in figure 7. The result showed that the TP rate which gave a description of the proportion of actual cases that was correctly predicted was 0.812 which implied that 81% of the actual cases were correctly predicted; the FP rate which gave a description of the proportion of actual cases misclassified was 0.105 which implied that 11% of actual cases were misclassified while the precision which gave a description of the proportion of predictions that were correctly classified was 0. 852 which implied that 85% of the predictions made by the classifier were correct. For the Multi-layer Perceptron algorithm based on the results presented in the confusion matrix presented in figure 9. The result showed that the TP rate which gave a description of the proportion of actual cases that was correctly predicted was 1, which implied that all of the actual cases were correctly predicted; the FP rate which gave a description of the proportion of actual cases misclassified was 0 which implied that none of actual cases were misclassified while the precision which gave a description of the proportion of predictions that were correctly classified was 1, implying that all of the predictions made by the classifier were correct.

In general, the multi-layer perceptron and the C4.5 decision trees algorithm were able to predict the risk of cataracts better than the naïve Bayes classifier algorithm. Although, the difference between the performance of the Naïve Bayes' classifier and the C4.5 Decision Trees classifier was 1 misclassification. Overall, the Multi-layer Perceptron was able to accurately classify all cases of cataracts with a value of 100% showing that it had the capacity to identify the complex patterns that existed within the dataset than the Naive Bayes classifier and the C4.5 decision trees algorithm. The variables identified by the C4.5 Decision Trees algorithm can also be given very close attention and observation, in order to better understand the risk of cataracts in patients monitored by the ophthalmologists.

5. CONCLUSIONS

This study focused on the development of a prediction model using identified risk factors in order to classify the risk of cataracts in selected respondents for this study. Historical dataset on the distribution of the risk of cataracts among respondents was collected using questionnaires following the identification of associated risk factors of cataracts from expert ophthalmologists. The dataset containing information about the risk factors identified and collected from the respondents was used to formulate predictive models for the risk of cataracts using C4.5 decision trees, naïve Bayes' and multi-layer perceptron classifier algorithms. The predictive model development using the algorithms were formulated and simulated using the WEKA software.

The results of the study revealed the variables that were identified by the C4.5 Decision Trees algorithm to be predictive for the risk of cataracts. Following the comparison of the performance of the machine learning algorithms used in this study, it was observed that the Multi-layer Perceptron had the best capability to identify the unseen patterns existing within the variables used to formulate the predictive model for the risk of cataracts. Following the development of the prediction model for cataracts' risk classification, a better understanding of the relationship between the attributes relevant to cataracts risk 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 cataracts risk classification prediction model, thereby improving the clinical decisions affecting cataracts' risk and the real-time assessment of clinical information affecting cataracts' risk from remote locations. It is advised that a continual assessment of variables monitored for cataracts' risk be made in order to increase the number of information relevant to creating an improved prediction model for cataracts' risk classification using the proposed model in this study.

REFERENCES

[1] World Health Organization (WHO). (2014). Visual impairment and blindness. Fact Sheet No. 282. Accessed from http://www.who.int/mediacentre/factsheets/fs282/en/ on July 12, 2017.

[2] Pascolini, D. and Mariotti, S.P. (2010) Global estimates of visual impairment. *British Journal of Ophthalmology 96*(5): 614 – 618. Accessed from http://dx.doi.org/10.1136/bjophthalmol-2011- 300539 on July 12, 2017.

[3] Bradford C (2004). *Basic Ophthalmology 8th Edition*. American Academy of Ophthalmology: 7-16

[4] Ajaiyeoba, A.I. and Fasina, F.O. (2003). The prevalence and cause of blindness and low vision in Ogun state. *Nigerian African Journal of Biomedical Research* 6(2): 63 – 67.

[5] Raizada, I.N., Mathur, A. and Narang, S.K. (1984). A study of prevalence and risk factors of senile cataract in rural areas of Western U.P. *Indian Journal of Ophthalmology 32*(5): 339 - 342

[6] Sobti, S. and Sahni, B. (2013). Cataract among Adults aged 40 years and above in a rural area of Jammu district in India: Prevalence and Risk Factors. *International Journal of Healthcare and Biomedical Research 1*(4): 284 – 296.

[7] Kyari, F., Gudlavalleti, M.V., Sivsubramanian, S., Gilbert, C.E., Abdull, M.M. and Entemekume, G. (2009). Prevalence of blindness and visual impairment in Nigeria: The National Blindness and Visual Impairment Study. *Investigating Ophthalmological Visual Science 50*(5): 2033 - 2039.

[8] Isaacs, R., Ram, J. and Apple, D. (2004). Cataract blindness in the developing world: is there a solution? *Journal of Agro-Medicine 9*: 207 – 220.

[9] Abdul, M.M., Sisvasubramanian, S., Murthy, G.V., Gilbert, C., Abubakar, T. and Ezelum, C. (2009). Causes of blindness and visual impairment in Nigeria: The Nigeria National Blindness and Visual Impairment Survey. *Invest. Ophthalmology. Vis. Sci. 50*(9): 4114 - 4120.

[10] 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.

[11] Waijee, A., Mukherjee, A. and Singal, A. (2013). Comparison of modern imputation methods for missing laboratory data in medicine. BMJ Open 3(8): 1 – 7.

[12] Bigus, J.P. (1996). *Data Mining with Neural Networks*. New York: McGraw- Hill.

[13] Hemalatha, M. and Megala, S. (2011). Mining Techniques in Healthcare: A Survey of Immunization. Journal of Theoretical and Applied Information Technology 25(2): 63 – 70.

[14] 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.

[15] Bowd, C. and Goldbaum, M. (2008). Machine learning Classifiers in Glaucoma. *Journal of Optometry and Vision Science 85*(6): 396 – 405.

[16] Imberman, S.P., Ludwig, I. and Zelikovitz, S. (2011). Using Decision Trees to Find Patterns in an Ophthalmology Dataset. In *Proceedings of the 24th International Florida Artificial Intelligence Research Society Conference 12*: 3 – 14.

[17] Kabari, L.G. and Nwachukwu, E.O. (2012). Neural Networks and Decision Trees for Eye Diseases Diagnosis. In *Advances in Expert Systems*. Accessed from http://dx.doi.org/10.5772/51380 on July 12, 2017.

[18] 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* (IRJET) 3(3): 547-551.