Development of a Risk-Factor Model for Predicting Occurrence of Knee Osteoarthritis

[1]Oladejo R.A. & [2]Achori B.T.
Department of Computer Sciences
School of Pure and Applied Science
Ogun State Institute of Technology
Igbesa , Ogun State, Nigeria
Email: funrack2002@gmail.com
Phone+234839722471

ABSTRACT

This study identified the required risk factors for Knee Osteoarthritis (KOA) patients and formulated a predictive model based on the identified variables. Extensive review of related work was done so as to understand the body of knowledge surrounding musculoskeletal related diseases and to identify knee osteoarthritis as one of the diseases under musculoskeletal condition as well as elicit the risk factors for it, these were validated from medical experts. The model to forecast knee osteoarthritis was formulated comparing four supervised machine learning algorithms namely Naïve Bayes, Multilayer perceptron, C4.5 Decision Tree and Support Vector Machine. The result of the model showed an accuracy of 97.59% considering the 36 initially identified attributes using no feature selection method, the results also showed the minimum number of variables relevant for knee osteoarthritis condition. Further results showed that all identified variables are relevant for effective and efficient development of a prognostic model for knee osteoarthritis. The study concluded that age as the most important variable for KOA and that all 36 identified attributes are relevant for predicting the risk of KOA.

Keywords: Knee osteoarthritis, Prognostic Model, Machine learning.

Proceedings Reference Format
Oladejo R.A. & Achori B.T.. (2021 Development of a Risk-Factor Model for Predicting Occurrence of Knee Osteoarthritis. Proceedings of the 27th iSTEAMS Multidisciplinary Innovations & Technology Transfer (MINTT) Conference. Academic City University College, Accra, Ghana. June, 2021. Pp 431-464 www.isteams.net/ghana2021. DOI - https://doi.org/ 10.22624/AIMS/iSTEAMS-2021/V27P40

1. INTRODUCTION

Health is a critical aspect of life. This simply means the absence of any illness, symptoms, morbidity and the capacity to attain one's own goals through target-oriented actions (World Health Organisation, 2001). Musculoskeletal disorder is an ill health related disease. It is a prevalent condition that affects the musculoskeletal system. This system includes; joints, ligaments, muscles, nerves, tendons, structures that support limbs, neck and back (Mirer and Stellman, 2008). The impact of musculoskeletal disorders is pervasive as they are a major burden on individuals alongside the health and social care systems (Woolf and Pfleger, 2003). In 2012, survey shows that musculoskeletal conditions ranked the second greatest cause of disability since they affect more than 1.7 billion people worldwide. In the United States alone, it ranks first among diseases using measures of disability (United State Bone Joint Initiative, 2015). There are more than 100 types of musculoskeletal disorder of which Osteoarthritis is one of them.

Osteoarthritis (OA), otherwise called degenerative joint disease or wear and tear disease is the most common musculoskeletal disorder. It is a chronic progressive disease that affects the cartilage (Kontzias, 2017; Kerkar, 2017). A cartilage is a firm, rubbery material that covers the end of each bone in a normal joint. It allows smooth gliding of the bones and serve as a cushion between the bones, However, this breaks down in OA infected joints, causing pain, swelling and the problem of moving the joint (Arthritis Foundation, 2017). As OA worsens over time, bits of bone or cartilage chip off. This floats around in the joint or developed into a growth called spurs. In the final stages of osteoarthritis, the cartilage wears away and bone rubs against bone, this leads to joint damage and more pain (Riviere, 2017). Inspite of billions of dollars expended on research, no drugs has been proven to modify the biological progression of OA so far, only a few treatment has been shown to relieve the symptoms beyond the placebo effect (Gardiner *et al.*, 2016). Therefore, there is a need to prevent this disease.

Presently, there is no cure or treatment that can reverse the damage of OA in the joints, and joint replacement is expensive, invasive and only effective for treating end-stage OA in older people (Longton *et al.*, 2016). Disease progression is usually slow but can ultimately lead to joint failure and disability. This disease is characterised by severe joint pain, tenderness, limitation of movement, stiffness of joint, occasional effusion, and variable degrees of local inflammation. There are more than 100 types of arthritis, but OA is the most common of all (Arthritis Foundation, 2017). OA occur in any joint, but is most common in the hip, knee, and the joints of the hand, foot, and spine (Symmons *et al.,* 2006).

The prevalence of osteoarthritis increases indefinitely with age, because the condition is not reversible. Advanced age, obesity, genetics, gender, bone density, trauma and a poor level of physical activity can lead to the onset and progression of osteoarthritis (Gabay and Clouse, 2016). The chances of having OA are common in persons at age 45 years and above. 50% of persons over the age of 65 years and almost all persons at 85years and above are infected (Yarnell and O'Reilly, 2013). Men among those less than 45years of age are often affected than women of the same age bracket, whereas women are affected more frequently among those aged 55 years and above (Yarnell and O'Reilly, 2013).The burden of OA is physical, psychological and socioeconomic. It can be associated with significant disability, such as a reduction in mobility and activities of daily living. The Psychological effects include distress, devalued self-worth and loneliness. Given the high frequency of OA in the population, its economic burden is large (Litwic *et al.,* 2013).

Global burden of knee osteoarthritis affects approximately 250 million people (Murray *et al.,* 2013). Worldwide estimates are that 9.6% of men and 18.0% of women over the age of 60 years have symptomatic OA. Approximately 80% of those with OA will have limitations in movement, and 25% cannot perform their major activities of daily life (World Health Organisation, 2012). Both knee and hip osteoarthritis is ranked 11th among 291 conditions that leads to disability globally. The increase in life expectancy and ageing populations are expected to make OA the fourth leading cause of disability by the year 2020 (Kraus *et al*., 2015). As the world's population continues to increase, there is estimation that OA will impact at least 130 million individuals around the globe by the year 2050 (Maiese, 2016).

Despite the overwhelming reports on the rising prevalence of musculoskeletal conditions, reports from Africa are lacking and underestimated. In 2006, some studies from Nigeria, Liberia, and South Africa were used in estimating the burden of rheumatoid arthritis in Africa (Symmons *et al*., 2015). This showed a high male to female ratio that was inconsistent with global trends and literature (Symmons *et al*., 2015).

Similarly, only one study from South Africa was used in estimating the burden for osteoarthritis in Africa, this highlights the paucity of data in Africa (Symmons *et al.,* 2006). In Nigeria, one out of every five adults aged 40 years and above has symptomatic knee osteoarthritis with a point prevalence of 19.6% (Akinpelu *et al.,* 2009*).* In medical research, preventive medicine is an important area and making predictions is an essential part of preventive medicine and even corrective medicine. Diagnosing patients, prognosticating patient health status and classification of biomedical signal patterns are examples of prediction (Ali *et al*., 2015).

Prognostic models are models that are used for forecasting, estimating or predicting the risk of something happening in the future. The main aim of prognostic modelling is to adequately model a problem domain in order to make forecast (Staal, 1999). Accurate prognostic models can inform patients and physicians about the future course of an illness or the risk of developing it, thereby guiding decisions on prevention (Waijee *et al.,* 2013). Machine learning has found great importance in the area of predictive modelling in medical research (Deo, 2015). Machine learning techniques can be broadly classified into supervised and unsupervised techniques. Supervised techniques involves matching a set of input records to one out of two or more target classes while the unsupervised techniques is used to create clusters or attribute relationships from raw, unlabelled or unclassified datasets (Mitchell, 1996). Supervised machine learning algorithms can be used in the development of classification or regression models.

In this study, GA was used to select the relevant features, used for the proposed model. Thereafter, the following four (4) supervised machine algorithm: Naïve Bayes, C4.5 Decision Trees classifier, Multi-Layer Perceptron-based and Support Vector Machine (SVM) were used in the research work. The performance was evaluated and the most accurate was used to formulate the prognostic model. Osteoarthritis typically develops over decades, offering a long window of time to potentially alter its course. Epidemiological and genetic studies of OA indicate that many pre-OA disease states can be modified (Chu e*t al*., 2012). The non-modifiable risk factors include gender and age whereas the modifiable risk factors include body mass index (BMI), injury/trauma, among others. The aetiology of OA is multifactorial, showing strong associations with highly modifiable risk factors of mechanical overload, obesity and joint injury (Lohmander, 2000).

KOA which is an age relevant disease is a major prevalent arthritic disorder globally. Despite extensive research costing billions of dollars, no drugs have been proven to modify the biological progression of it and only a few treatments are proven to relief the symptoms (Gardiner *et al*., 2016). Since treatment options for knee osteoarthritis (KOA) are limited, attention has been shifted to predicting the risk, this is critical to support paradigm shift from palliation of the disease to prevention. Presently, there is no model that considered requisite number of variables to predict KOA despite the fact that it is one of the leading cause of disability and immobility. There is a need to formulate a model to assist in estimating the risk of KOA with necessary and required predictors. This model when implemented can be integrated into existing health information systems thereby influencing real-time analysis of KOA clinical information. This will help relevant health care personnel, patients and other stakeholders to be able to make vital decisions and allocate resources such as materials, personnel and health services to areas deem necessary to combat prevalence.

2. REVIEW OF RELATED WORKS

2.1 Related Work

A number of works had been published in the area of application of machine learning algorithms to osteoarthritis risk prediction, classification and diagnosis. In risk prediction, none of the works published have considered menopause, good gait and climbing stairs as variables to predict the risk. A number of such works have however stressed the effect of machine learning in developing effective and efficient prediction models. Persson and Rietz (2017) applied machine learning algorithm for the prediction and analysing of osteoarthritis patient outcomes. Fifteen (15) variables were considered and comparative analyses were performed on five algorithms (Logistic regression, Random forest, Adaptive boosting, Gradient boosting and Multi-layer perceptron) without the application of feature selection algorithms to identify relevant features. Best result was obtained with Gradient Boosting model. Fewer variables were used and the following important variables were not considered; menopause, family history and gait.

Zhang *et al* (2011) worked on Nottingham knee osteoarthritis risk prediction models. A risk prediction models for knee osteoarthritis (OA) was developed, also an estimation of the risk reduction that results from modification of potential risk factors was made. Nine (9) variables were considered, only logistic regression model was used, some important variables such as Menopause, Leg deformation and Gait were not considered and also fewer variables were used.

Kumar *et al* (2017) applied machine learning to build a predictive model on knee osteoarthritis. A comparative analysis of two models Logistic regression (LR) and Naive Bayes (NB) was done to predict the estimated risk on a patient's chance of obtaining OA. Logistic Regression gives best fitting model for the prediction. This study is relevant due to the use of machine learning in predicting model on OA, however, Nine (9) variables were considered and some important variables such as Menopause, Leg deformation and Gait were not considered and just two algorithms were compared.

Black *et al* (2017) worked on framework for prognostic predictive model development using electronic medical record data with a Case Study in Osteoarthritis Risk. A Frame work for Prognostic model was developed, which outlines step-by-step guidance for the construction of a prognostic predictive model using EMR data. Logistic regression was used to predict osteoarthritis based on age, sex, BMI, previous leg injury, and osteoporosis. Few variable were considered and important variables like occupation, sport and leg deformation were not used. Most of the existing models were developed with few risk factors (some salient factors like sport, family history, leg deformation and so on were omitted). Many of these models are foreign having different environmental factors, nutrition, climates, occupation and accessibility to medical care, influencing the results,this make this paper to be quite different from all the existing Predictive model on KOA.

3. METHODS

In order to develop the predictive model to forecast the risk of KOA among individuals in Nigeria, a number of methods which include extensive review of related work to identify and elicit the risk factors for knee osteoarthritis was done. Five (5) physiotherapists were interviewed so as to validate the risk factors identified from literature. Relevant data containing the necessary risk factors required for monitoring knee osteoarthritis were collected from hospital medical records and questionnaires.

Genetic Algorithm was used for feature selection and thereafter the model to forecast knee osteoarthritis was formulated based on the use of supervised machine learning algorithms. The formulated model was simulated using the explorer Waikato Environment for Knowledge Analysis (WEKA) software. The historical data that was collected was used to validate the performance of the prognostic model by determining the true positive rate, false positive rate, accuracy and precision.

3.1 Data Identification and Collection

The first step in this study was to determine the variables that are associated with the risk of Knee Osteoarthritis (KOA). Thereafter, variable types were declared as Numeric, Nominal and Integer respectively depending on the variable name. Following the process of identifying the variables that are associated with the risk of KOA, a proforma was designed with the help experts to elicit data explaining the values of the indicators of the risk of KOA from a number of patients in the physiotherapy unit of OAUTHC while structured questionnaire was used to assess data from undiagnosed individuals in order to predictive the risk of KOA.

The first part of the questionnaire consists of the information required for identifying the patients selected considering their socio-demographic information as well as some clinical variables such as gender, occupation, ethnicity, age, state of origin, height, weight, body mass index (BMI) and so on. The demographic and clinical variables identified are shown in Table 1.

The second part of the questionnaire was used to collect information regarding the identified variables associated with the risk of KOA (see Table 2). Information regarding the variables associated with the risk of KOA includes painful episodes, Family history of KOA, sports engagement and social habits to mention a few. The final part of the questionnaire was used to ascertain the status of the individual respondent. That is, either respondent is KOA diagnosed or not based on the responses provided on the questionnaire. Other variables that were associated with the risk of KOA identified are shown in Table 2. Eighty three (83) data collected from a number of patients and healthy individuals were observed in the study.

Table 3.1: Demographic and Clinical Variables Identified

Variable Name	Variable Type	Values
Gender	Nominal	Male, Female
Age (in years)	Integer Numeric	
State of Origin	Nominal	36 states of the Federation
LGA of residence	Nominal	774 LGAs of the Federation
Ethnicity	Nominal	Yoruba, Hausa, Ibo, Others
Occupation	Nominal	Technician, Nurse, Business owner, Trader, Student, Trader, Retired, Farmer, Unemployed, Lecturer, Tailor, Civil Servant, Artisan, Teacher, Clergy, Engineer, Accountant, Manager, Clerk
Height (in m)	Real Numeric	
Weight (in Kg)	Real Numeric	
Body Mass Index (BMI)	Real Numeric	
BMI Classification	Nominal	Underweight, Normal, Overweight, Obese
Alcoholic	Nominal	Yes, No
Smoker	Nominal	Yes, No

Table 3.2: Other Associated Variables Identified

Variable Name	Variable Type	Values
Previous Injury	Nominal	Yes, No
Unequal Leg Length	Nominal	Yes, No
Family History	Nominal	Yes, No
Sport Engagement	Nominal	Yes, No
Pain (climbing staircase)	Nominal	Yes, No
Pain (walk long distance)	Nominal	Yes, No
Pain (load bearing)	Nominal	Yes, No
Pain (walking)	Nominal	Yes, No
Pain (when rested/ sleeping)	Nominal	Yes, No
Pain (joint pressed)	Nominal	Yes, No
Visible Swell on joints	Nominal	Yes, No
Stiff Joints	Nominal	Yes, No
Warmness on joints	Nominal	Yes, No
Crackling sound when walking	Nominal	Yes, No
Diabetic	Nominal	Yes, No
Menopause	Nominal	Yes, No, NA
Prostate gland	Nominal	Yes, No, NA
Leg deformation	Nominal	Yes, No, Not sure
Hypertensive	Nominal	Yes, No
Depression	Nominal	Yes, No
Good gait	Nominal	Yes, No

3.2 Data Pre-processing

Data pre-processing involved the process of data cleaning which ensures that all variables were properly identified with their pre-defined values as presented in Tables 1 and 2. All variables that were incorrectly represented were corrected while missing values were ignored. Missing variables which can be assessed from other variables were replaced with estimated values. Also, the BMI calculated from the height and weight of the patient were classified into 4 nominal values, each represented a valid interval of BMI values.

The data were stored in a comma separated variable (csv) file format which places the attributes collected in the first line and separated by a comma followed by data values for each patient record on each successive row. The last variable defined for each record presented were the risk of KOA associated with the information about the variables provided.

3.2.1 Feature selection using genetic algorithm

The motive of the feature selection algorithm was to be able to select *r* number of variables which may be more relevant to determining the risk of KOA from the *i* initially identified variables, where $r < i$. For the purpose of this study, G.A, a meta-heuristic computational intelligence algorithm was applied for the selection of relevant variables from the initially identified variables. The selected *r* attributes performance may have equal or greater performance than using the *i* initially identified variables according to equation (1) such that $X_r \subset X_i$.

$$Performance[f(X_i)] \leq Performance[f\{X_r)] \quad (1)$$

Therefore, the genetic algorithm (GA) chosen for this study was used to transform the dataset with *n* records and *i* attribute into a dataset with *n* records and *r* attributes where $r < i$ according to equation (2). $X^{n\times i}$ is a data matrix containing *n* records with *i* initially identified variables while $X^{n\times r}$ is a data matrix containing *n* records with *r* reduced attributes and $i, n \; and \; r \; \in Z^+$.

$$FS_{GA}: X^{n\times i} \longmapsto X^{n\times r} \quad (2)$$

The GA; a population-based search heuristic algorithm, which mimics natural evolution process, was used to extract the most relevant variables from the initially identified set of variables in this study. The GA employed the use of one population of chromosomes (called the solution candidates) for getting a new population by using a method of natural selection combined with mutation and crossover techniques. In comparison to human genetics, chromosomes are the bit strings (set of attributes selected (1) or rejected (0), gene is the attribute, allele is the attribute value (value in cell), locus is the bit position (attribute of interest), and genotype is the encoded bit string while phenotype is the decoded genotype. The fitness of the chromosomes was evaluated using an objective function or fitness function.

Figure 1 shows the flowchart of the genetic algorithm used for the extraction of relevant features using the process of chromosome encoding, population initialization, fitness evaluation, selection followed by genetic operators (mutation and/or crossover). The GA operated on a binary search space, the chromosomes are bit strings representing the set of attributes selected or ignored. The process of feature selection of relevant attributes began with the selection of an initial random population of attributes. Thereafter; the fitness was evaluated using the fitness function.

According to this study, a gene value of "1" indicates that a particular attribute indexed by the position "1" is selected, while a gene value of "0" indicates that an attribute indexed by position "0" is selected. The value of the fitness of chromosomes selected can be evaluated using a fitness function as shown in equation (3). The values α represent the error in classification using a KNN classifier and N_f as the number of attributes selected.

$$Fitness\ Function = \frac{\alpha}{N_f} + e^{\left(-\frac{1}{N_f}\right)} \qquad (3)$$

After the elite individuals are moved to the next generation, the remaining individuals in the current population were used to produce the rest of the next generation through crossover and mutation. Crossover involved the combination of two individual chromosome bit strings using modulo 2 arithmetic additions as defined in equation (4) to form a single chromosome bit string

$$cross - over = \ ChromBitString1 \oplus_{mod\ 2} ChromBitString2 \qquad (4)$$

Mutation on the other hand, was performed by flipping the bit strings based on the probability of mutation provided to the GA used. The surviving chromosome from the GA process were the best chromosome such that the index of the bit string for which there was a value of "1" identified the positions of the most relevant features selected by the GA. The parameters that were used to implement the GA proposed for feature selection in this study is presented in Table 3. The results of the application of GA on the dataset containing the initially identified variables reduced the datasets to one containing the relevant attributes associated with KOA.

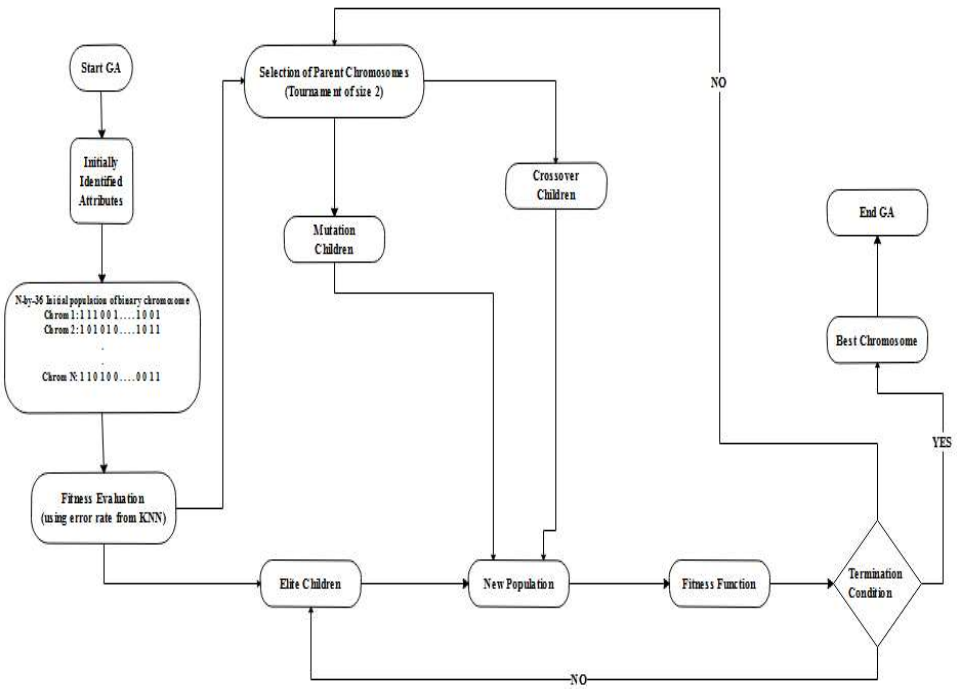

Figure 1: Flowchart of Genetic Algorithm for Feature Selection

3.3 Model Formulation

For the purpose of this study, supervised classification techniques was applied for the development of the prognostic model required for the risk of KOA using the identified variables associated with the risk of KOA. Since the task of the study is to be able to assess the presence or absence of the risk of KOA, then the required task is a classification problem aimed at identifying the results of the values of the set of attributes provided as an output for defining the respective risk of KOA.

The supervised machine learning algorithms that were developed are required to provide a mapping of the values of the set of input attribute set *X* to a target class set *Y* consisting of Yes or No cases as presented by equation (5).

$$f(X_i) = f(X_1, X_2, X_3 \ldots\ldots, X_i) = \{Yes\ No \quad (5)$$

Supervised machine learning algorithms are black-boxed models since it is not possible to give an exact description of the mathematical relationship existing between the input attribute set and the risk of knee osteoarthritis. However, cost functions were used by the supervised machine learning algorithms to estimate the error in prediction during the process of using the training data required for model development. The supervised machine learning algorithms selected for this study composes of algorithms that are stochastic and perceptron-based. The stochastic algorithms that were applied are the Naïve Bayes and the C4.5 decision trees algorithm while the perceptron-based algorithms used were the support vector machines and the multi-layer perceptron

Table 3.3: Parameters of GA Used for Feature Selection

GA Parameter	Value
Population Size	200
Genome Length	20
Population Type	Bit string of length 36
Fitness Function	KNN
Number of Generations	100
Crossover	Modulo 2 Addition
Crossover Probability	0.8
Mutation	Uniform Mutation
Mutation Probability	0.1
Selection Scheme	Tournament size of 2
Elite Count	2

.

3.3.1 Stochastic-based supervised machine learning algorithms

For this study, the naïve Bayes and the C4.5 decision trees classifiers which are stochastic-based Machine Learning (ML) algorithms were considered for the formulation of a prognostic model for the risk of knee osteoarthritis. The algorithms are presented in the following paragraphs.

a. Naïve Bayes' (NB) classifier for the risk KOA

The naive Bayes' Classifier is a probabilistic model based on the Bayes' theorem of conditional probability and is one of the most frequently used methods for supervised learning. It provides an efficient way of handling any number of attributes or classes which is 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 attributes associated with the risk of knee osteoarthritis alongside their respective risk of knee osteoarthritis, *C* (target class) collected for *j* number of records/patients and $H_k = \{H_1 = Yes, H_2 = No\}$ be a hypothesis that X_{ij} belongs to class C. For the classification of the risk of knee osteoarthritis given the values of the risk factor of the jth record, Naïve Bayes' classification required the determination of the following:

- $P(H_k|X_{ij})$ – Posteriori probability: is the probability that the hypothesis, H_k holds given the observed data sample X_{ij} for $1 \leq k \leq 2$.
- $P(H_k)$ - Prior probability: is the initial probability of the target class $1 \leq k \leq 2$;
- $P(X_{ij})$ is the probability that the sample data is observed for each risk factor (or attribute), *i*; and
- $P(|X_{ij}|H_k)$ 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 hypothesis H_k is defined according to Bayes' theorem as shown in equation (6) for each class. The risk of knee osteoarthritis is then determined by equation (7) based on the outcome for each class in equation (6)

$$P(X_{ij}) = \frac{\prod_{i=1}^{n} P(X_{ij}|H_k)P(X_i)}{P(H_k)} \quad for\ k = 1,2 \tag{6}$$

$$Risk\ of\ Osteoarthritis max.\left[P(X_{ij}), P(X_{ij})\right] \tag{7}$$

b. C4.5 Decision Trees classifier for the risk of KOA

The decision trees is a supervised machine learning algorithm which applies a divide-and-conquer approach for the growth of a recursive hierarchical tree which can be interpreted as a set of If-Then statement or rules which combines the attributes in such a way that it is possible to determine the risk of knee osteoarthritis. In order to do this, the pattern in the dataset was learnt by the tree by splitting the training dataset into subsets based on an attribute value test for each input variables; the process was then 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 also called the Top-down induction of trees.

In theory, the 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 (risk of knee osteoarthritis) given the variables represented by the path from the root node. Rules were then induced from the trees taking paths created from the root node all the way to their respective leaf using IF-THEN statements. The decision trees algorithm was then required to distinguish between important variables attributes and attributes which contribute little to overall decision process which are based on the use of impurity measures.

Figure 3 shows the algorithm used by decision trees in growing trees from a dataset containing a set of attributes. The algorithm is called Tree Growth and takes in two arguments; which are the training records containing the records *E* and the attribute set (variables associated with risk of knee osteoarthritis) F which works by recursively splitting the data and expanding leaf nodes until a stopping criterion is met. The stopping criteria that was used by the C4.5 decision tree is the Gain ratio which determines the information gain of each attribute shown in equation (8) at every node generation and divides it by the split value according to equation (10).

Therefore, the attribute X_i with the greatest value of the gain ratio was then chosen

$$IG(X_i) = H(X_i) - \sum_{t \epsilon T} \frac{|t|}{|X_{ij}|} \cdot H(X_i) \quad (8)$$

Where: $$H(X_i) = -\sum_{t \epsilon T} \frac{|t,X_i|}{|X_{ij}|} \cdot \frac{|t,X_i|}{|X_{ij}|} \quad (9)$$

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 Osteoarthritis Risk class to leaf label
Return leaf
else
root = creatNode( )// 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 split
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 (attributes)
end for
end if
return root

Figure 3: C4.5 Decision Trees Algorithm

$$Split(T) = -\sum_{t \epsilon T} \frac{|t|}{|X_{ij}|} \cdot \frac{|t|}{|X_{ij}|} \quad (10)$$

as a potential node and its value

$$\sum_{k=1}^{i} w_k x_k = w_1 x_1 + w_2 x_2 + \cdots + w_i x_i =< w.x > \quad (11)$$

s $t \in T$ was used to split the dataset after which subsequent attributes were determined for splitting the trees till the terminal nodes was reached.

3.3.2 Perceptron-based Supervise Machine Learning Algorithm

Perceptron-based ML Algorithm are classifiers which required inputs to be fired as neurons via synaptic weights assigned such that the input is the sum of products of the weights w_{ij} attached to each input X_i at a node *j* as shown in equation (11). For this study, Support Vector Machine and Multilayer Perceptron (MLP) were considered for formulating a prognostic model for the risk of knee osteoarthritis

a. Support Vector Machines (SVM) for the risk of KOA

An SVM model is a representation of the dataset as points in space, mapped so that the datasets of the separate categories will be divided by a clear gap that is as wide as possible. New examples will then be mapped into that same space and will be used to predict the risk class category of the example based on which side of the gap they fall on. In formal terms, SVM was used to construct a hyper-plane in a high-dimensional space, which was applied for the classification of the dataset. A good separation was achieved by the hyperplane that had the largest distance to the nearest training data points X_i which is called the support vectors since in general the larger the margin, the lower the generalization error of the classifier.

Therefore, the SVM during model formulation was attempted to minimize the cost of classification by maximizing the distance between hyper-planes. A good separation was achieved by the hyperplane $< w, x > +b = 0$ that had the largest distance $\frac{2}{||w||}$ to the neighbouring data points of either classes at opposite ends, since in general the larger the margin the lower the generalization error of the SVM classifier. Figure 4 shows the separation of the different two (2) risk classes of knee osteoarthritis from the dataset. It shows how the hyper-plane was used to separate the dataset into two (2) risk classes, such that hyperplane was used to differentiate between the cases with risk of knee osteoarthritis and those without the risk of knee osteoarthritis cases. On the other hand, Figure 5 shows a clear description of the relationship between the SVM parameters and the hyper-plane used in separating the margins from the support vectors X_i.

The hyperplane created was defined as $< w, x > +b = 0$ where $w \epsilon R^p$ and $b \epsilon R$ while $< w, x > +b = -1$ and $< w, x > +b = 1$ are the margins required for the separation *w* of support vectors *x* within the *n* variables. Therefore, equation (12) was used to defined a linearly separable function for which the decision function in equation (13) was used to propagate the output of equation (12) using a sigmoid function with interval {-1, 1}.

The aim of the SVM is to maximize the separation of the hyper-planes in equation (13) subject to the decision function defined in equation (14).

$$Risk_i = f(x_i) = (< w, x_i > + b) > 0, \quad \forall i \in [1, n] \tag{12}$$

$$f_d(x_i) = sign(Risk_i) = (< w, x_i > + b) > 0, \quad \forall i \in [1, n] \tag{13}$$

$$maximize \ \frac{1}{2}||w||^2 \tag{14}$$

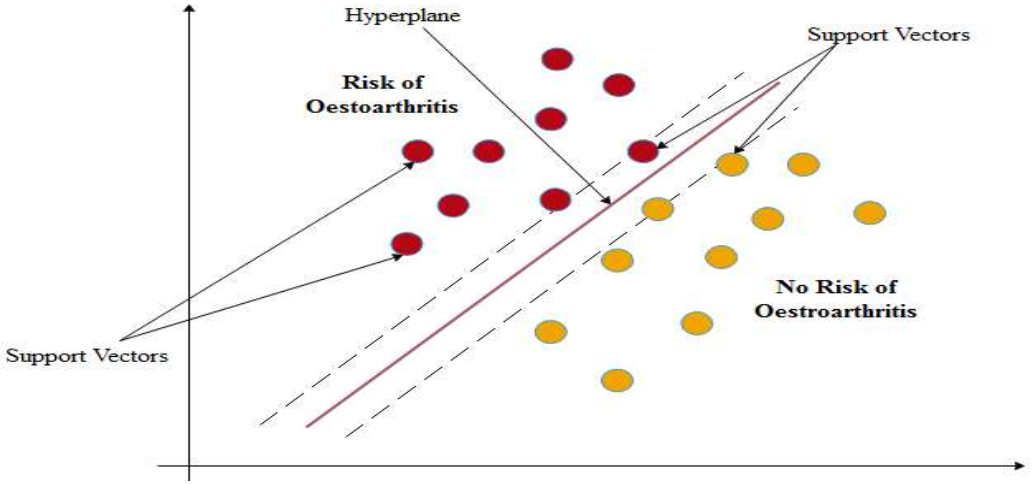

Figure 4: The Separation of the Datasets using SVM Hyperplane

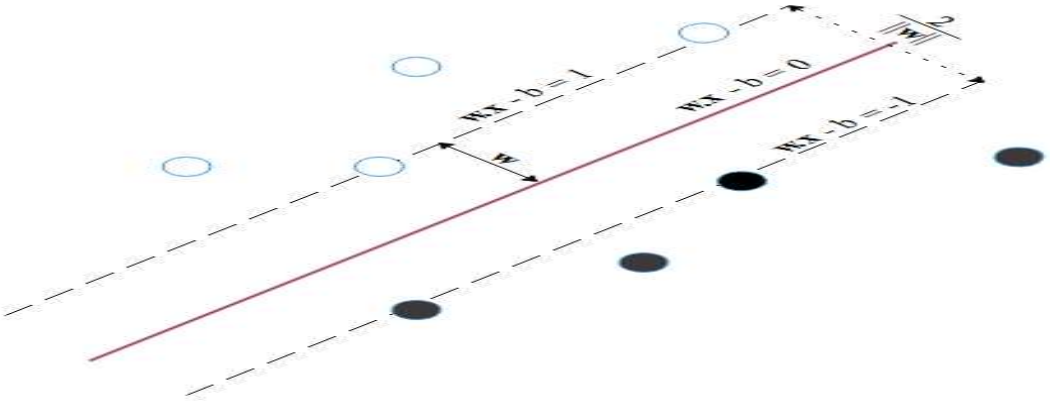

Figure 5: SVM Hyperplane showing the underlying parameters

b. Artificial Neural Network – Multi-layer perceptron (MLP) for the risk of KOA

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;

$$output\ O_j = \varphi(\sum_{k=1}^{i} w_{kj}x_k + b_k) = \varphi(z) = \frac{1}{1+e^{-z}} \quad (15)$$

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.

$$\delta_j = \frac{\partial E}{\partial O_j}\frac{\partial O_j}{\partial net_j} = \{(O_j - p_j)\varphi(net_j)\big(1 - \varphi(net_j)\big)\ j\ is\ output\ neuron, (\sum_{l\epsilon L} \delta_j w_{jl})\varphi(net_j)\big(1 - \varphi(net_j)\big)\ j\ is\ hidden\ neuron \quad (16)$$

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

$$\frac{\partial E}{\partial w_{ij}} = \delta_j x_i \quad (17)$$

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

$$\Delta w_{ij} = -\alpha \frac{\partial E}{\partial w_{ij}} \quad (18)$$

3.4 Performance evaluation metrics

In order to evaluate the performance of the supervised machine learning algorithms used for the classification of the risk of knee osteoarthritis, there was a need to plot the results of the classification on a confusion matrix as shown in Figure 7. All correct classifications lie along the diagonal from the north-west corner to the south-east corner also called True Positives (TP) and True Negatives (TN) while other cells are called the False Positives (FP) and False Negatives (FN). the likely cases are considered as the positive case while the unlikely and probable cases are the negative cases; the definitions are presented such that the True positives (TP) are correctly classified Yes cases, False positives (FP) are incorrectly classified No cases, True negatives (TN) are correctly classified No cases while False negatives (FN) are incorrectly classified Yes cases.

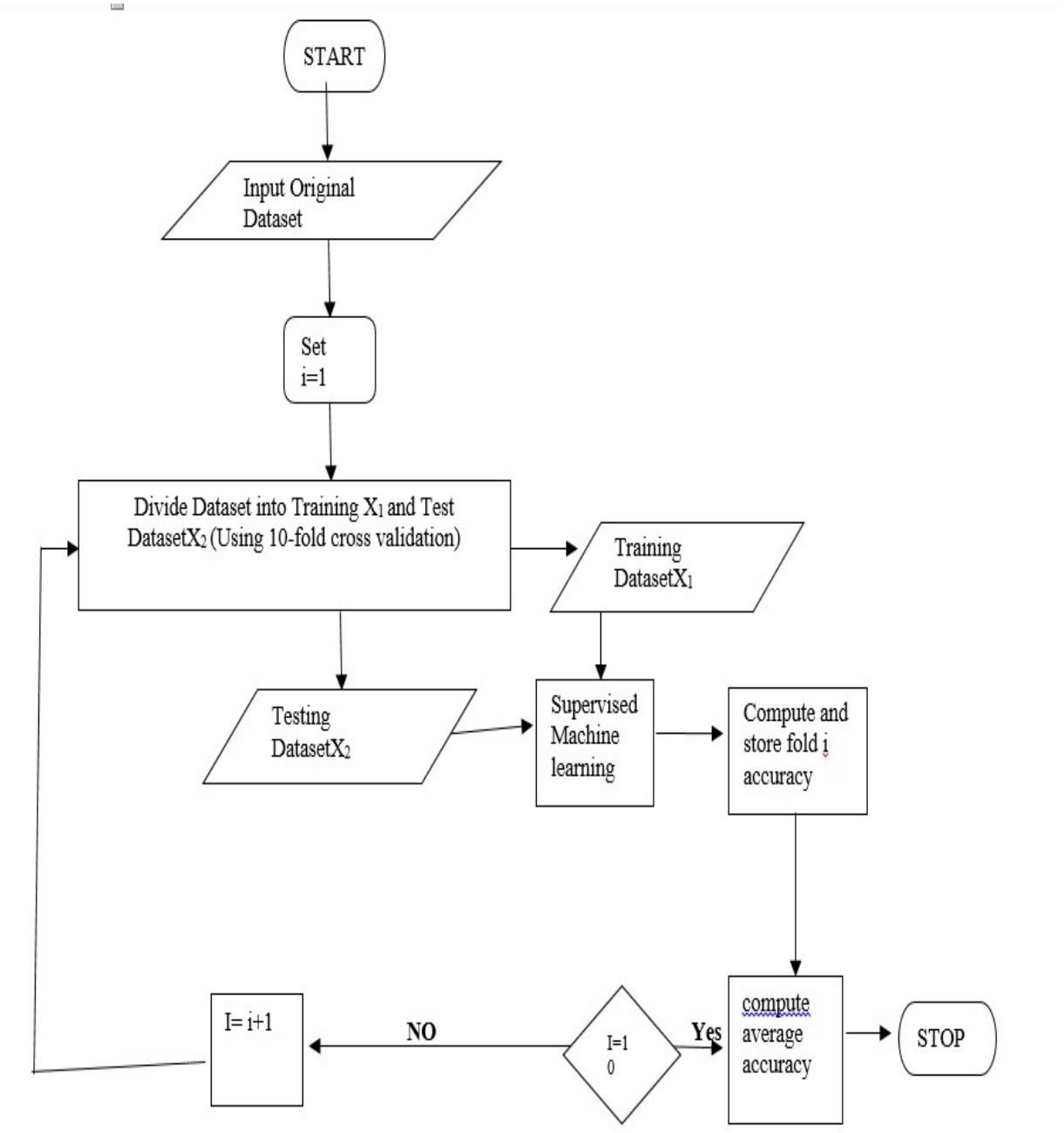

Figure 6: Flowchart of the 10-fold Cross Validation Process for Training

The true positive/negative and false positive/negative values recorded from the confusion matrix were then used to evaluate the performance of the prediction model based on a number of performance evaluation metrics.

A description of the definition and expressions of the metrics are presented as follows:

a. **True Positive (TP) rates (sensitivity/recall)** – proportion of positive cases correctly classified.

$$TP\ rate_{Yes} = \frac{TP}{TP + FN} \quad (19)$$

$$TP\ rate_{No} = \frac{TN}{FP + TN} \quad (20)$$

	Predicted Yes	Predicted No
	TP	FN
	FP	TN
	Actual Yes	Actual No

Figure 7: Diagram of a Confusion Matrix

TP = Correctly classified Yes
FP = Incorrectly classified No
TN = correctly classified No
FN= incorrectly classified Yes
Actual Cases = No= TN+TP
Yes = TP+FN
Predicted Cases = No = FN+TN
Yes=TP+FP
False Positive (FP) rates (1-specificity/false alarms) – proportion of negative cases incorrectly classified as positives.

$$FP\ rate_{Yes} = \frac{FP}{F + TN} \quad (21)$$

$$FP\ rate_{No} = \frac{FN}{TP + FN} \quad (22)$$

a. **Precision** – proportion of predicted positive/negative cases that are correct.

$$Precision_{Yes} = \frac{TP}{TP + FN} \quad (23)$$

$$Precision_{No} = \frac{TN}{TN + FP} \quad (24)$$

b. **Accuracy** – proportion of the total predictions that was correct.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (25)$$

Using the aforementioned performance metrics, the performance of the prognostic model for the classification of risk of knee osteoarthritis was evaluated by validation using a historical dataset collected based on the information provided in the questionnaire. The TP rate and precision lie within the interval [0, 1], accuracy within the interval of [0, 100] % while the FP rate lies within an interval of [0, 1]. The closer the accuracy is to 100% the better the model, the closer the value of the TP rate and precision is to 1 the better while the closer the value of FP rate is to 0 the better. Therefore, the evaluation of an effective model has a high TP/Precision rates and a low FP rates.

4. RESULTS

4.1 Results of Data Identification and Collection

For this study, extensive review of literatures and consultations with domain experts was done to identify and elicit the risk factor variables for KOA. Thirty six (36) variables were identified and a total of 100 KOA records were collected from hospital records; (OAUTH Complex) and questionnaires. The data collected was studied in order to understand the pattern. Out of a total of 100 questionnaires that were administered, 83 questionnaires were filled by the patients and respondents, these responses were analysed using descriptive statistical frequency distribution tables in order to observe the distribution of the risk of knee osteoarthritis among the patients selected for this study. From the data collected, it was observed that 46 (55.4%) records consisted of patients with risk of knee osteoarthritis while 37 (44.6%) consisted of patients without the risk of knee osteoarthritis.

Table 4.1: Results of the Description of Demographic Variables

Variable Name	Values	Frequency (%)
Gender	Male	43 (51.8)
	Female	40 (48.2)
State of Origin	Osun	40 (51.8)
	Oyo	18 (21.7)
	Ogun	9 (10.8)
	Ondo	3 (3.6)
	Ekiti	3 (3.6)
	Others	10 (12.0)
Ethnicity	Yoruba	77 (92.8)
	Ibo	4 (4.8)
	Others	1 (1.2)
	Missing	1 (1.2)
Occupation	Technician	2 (2.4)
	Nurse	1 (1.2)
	Business owners	3 (3.6)
	Trading	12 (14.2)
	Student	8 (9.6)
	Retiree	15 (18.1)
	Farmer	3 (3.6)
	Unemployed	1 (1.2)
	Lecturing	5 (6.0)
	Tailor	1 (1.2)

	Civil-Servant	10 (12.0)
	Artisan	1 (1.2)
	Teaching	7 (8.4)
	Clergy	2 (2.4)
	Engineer	6 (7.2)
	Accountant	1 (1.2)
	Manager	3 (3.6)
	Clerk	2 (2.4)
BMI Classification	Underweight	1 (1.2)
	Normal	16 (19.3)
	Overweight	29 (40.0)
	Obese	27 (32.5)
	Missing	10 (12.0)
Alcoholic	Yes	3 (3.6)
	No	79 (95.2)
	Missing	1 (1.2)
Smoker	Yes	0 (0.0)
	No	82 (98.8)
	Missing	1 (1.2)

Table 4.2: Results of the Summary Statistics of the Numeric Variables

Variable	Minimum	Maximum	Mean	Standard Deviation
Age (in Years)	19.0	86.0	49.01	18.658
Weight (Kg)	52.0	109.0	73.70	12.058
Height (metres)	1.2	2.2	1.63	0.126
Body Mass Index (BMI)	16.0	40.9	28.74	5.012

Table 4.3: Results of the Description of the Associated Variables Identified

Variable Name	Values	Frequency (%)
Previous Injury	Yes	22 (26.5)
	No	57 (68.7)
	Missing	4 (4.8)
Unequal Leg Length	Yes	1 (1.2)
	No	81 (97.6)
	Missing	1 (1.2)
Family History	Yes	9 (10.8)
	No	71 (85.5)
	Missing	3 (3.6)
Sport Engagement	Regularly	14 (16.9)
	Seldom	50 (60.2)
	Not at all	17 (20.5)
	Missing	2 (2.4)
Pain (climbing staircase)	Yes	32 (38.6)
	No	49 (59.0)
	Missing	2 (2.4)
Pain (walk long distance)	Yes	26 (31.3)
	No	55 (66.3)
	Missing	2 (2.4)
Pain (load bearing)	Yes	43 (51.8)
	No	40 (48.2)
Pain (walking)	Yes	45 (54.2)
	No	37 (44.6)
	Missing	1 (1.2)
Pain (when rested/ sleeping)	Yes	23 (27.7)
	No	59 (71.1)
	Missing	1 (1.2)
Pain (joint pressed)	Yes	52 (62.7)
	No	31 (37.3)
Visible Swell on joints	Yes	22 (26.5)
	No	58 (69.9)
	Missing	3 (3.6)

Variable Name	Values	Frequency (%)
Stiff Joints	Yes No Missing	35 (42.2) 45 (54.2) 3 (3.6)
Variable Name	**Values**	**Frequency (100%)**
Warmness on joints	Yes No Missing	15 (18.1) 65 (78.3) 3 (3.6)
Crackling sound when walking	Yes No Missing	9 (10.8) 72 (86.7) 2 (2.4)
Diabetic	Yes No Missing	6 (7.2) 76 (91.6) 1 (1.2)
Menopause	Yes No NA	23 (27.8) 17 (20.5) 43 (51.9)
Prostate gland	Yes No NA Missing	4 (4.8) 38 (45.8) 40 (48.2) 1 (1.2)
Leg deformation	Yes No Not sure	10 (12.0) 70 (84.3) 3 (3.6)
Hypertensive	Yes No Missing	26 (31.3) 55 (66.3) 2 (2.4)
Depression	Yes No Missing	4 (4.8) 74 (89.2) 5 (6.0)
Good gait	Yes No	60 (72.3) 23 (27.7)

Majority of the patients were also observed to be non-diabetic (92%), non-hypertensive (66%), non-depressive (89%) and without leg deformation (84%). Figure 1 shows the screenshot of the distribution of the nominal and numeric features assessed for the risk of KOA for those with risk (blue) and those with no risk (red). Figure 2 shows the distribution of cases with risk and without risk of KOA using blue and red colours respectively. After the description of the identified variables that were associated with the risk of KOA, the process of the identification of the most relevant features using genetic algorithm was done and compared with the variables selected using traditional filter-based feature selection algorithms.

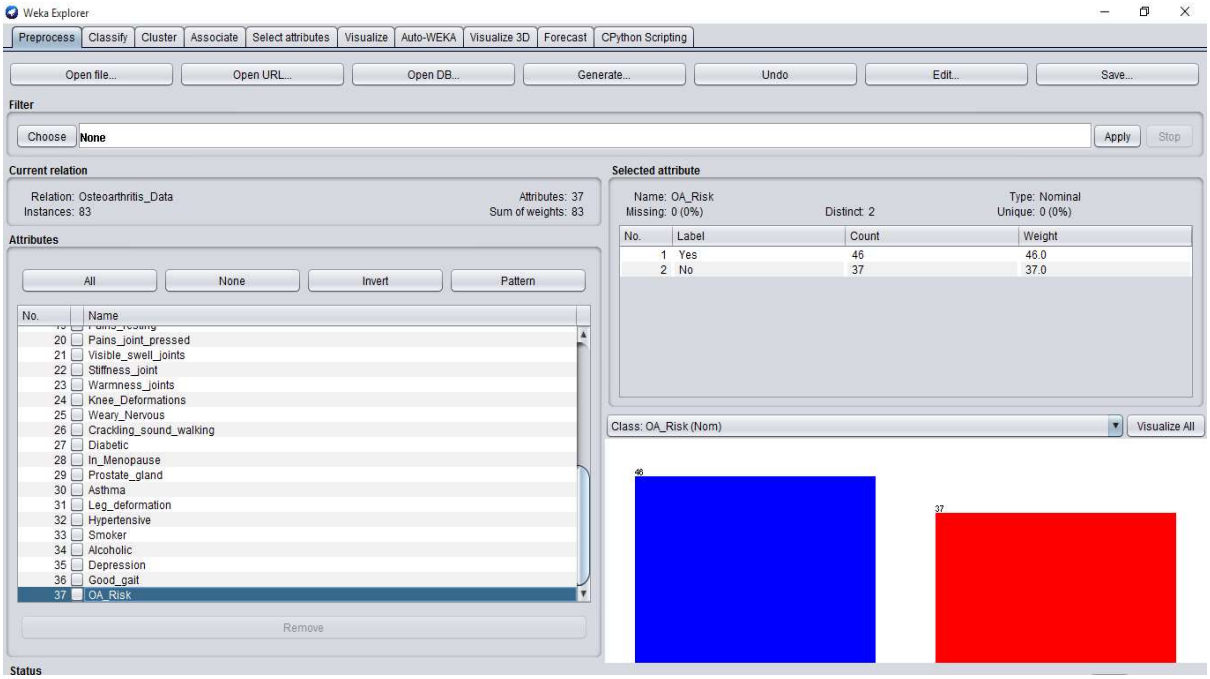

Figure 4.1: Screenshot of distribution of the nominal and numeric features assessed for the risk of KOA with risk With no risk

4.2.Consistency base Feature Selection (CFS)

In addition to the variables selected by the Genetic Algorithm, another feature selection algorithm was also applied for the identification of relevant variable. The subset evaluator such as the consistency based feature selection algorithm which ranked the attributes in the order of importance was applied.

4.3 Discussion of the Feature Selection results

Results of the relevant attributes selected by the genetic algorithm and the consistency based feature selection algorithm alongside the initially identified variables are presented in Table 4.4 below. The results showed that GA selected 12 relevant attributes out of the 36 initially identified variables while the consistency-based FS algorithms selected seven 7 relevant attributes. Out of the variables selected by the GA, the 7 variables selected by the consistency-based FS algorithms were also among them. Therefore, the GA was able to identify a larger number of variables than those that were selected by the consistency-based FS algorithms. Variables common to the two feature selection used are age, LGA of residence, pains (while climbing staircase), weight, pains (when joints are pressed), visible swelling on joints and good gait.

Variables that were not included in the feature selected but are among the 36 initially identified variables are previous injury, unequal leg length, family history, sport, some pain episodes, diabetic, depression and so on.The set of attributes identified by each FS algorithm alongside the initially identified attributes were sent to the supervised machine learning algorithms for estimating the performance of the prognostic models. Thereafter, the performance of the model developed using the relevant features selected and the 36 variables were compared to determine the best.

Table 4.4: Relevant Features selected by Genetic and Consistency FS Algorithms

Genetic Algorithm	Consistency-Based FS	Initially identified attributes
Age (in years)	Age (in years)	Gender
LGA of residence	LGA of residence	Age (in years)
Pains (while climbing staircase)	Pains (while climbing staircase)	State of Origin
Weight (Kg)		Weight (Kg)
		LGA of residence
Pains (while walking)	Pains (when joints are pressed)	Ethnicity
Pains (when joints are pressed)		Visible swelling on joints
		Occupation
Visible swelling on joints	Good gait	Height (in m)
Warmness on joints		Weight (in kg)
Feeling weary or nervous		BMI
Menopause		BMI Classification
Leg Deformation		Alcohol
Good gait		Smoking
		Previous Injury
		Unequal Leg Length
		Family History
		Sport Engagement
		Pain (climbing staircase)
		Pain (walk long distance)
		Pain (load bearing)
		Pain (Walking)
		Pain (when rested/sleeping)
		Pain (joint pressed)

	Visible swell on joints Stiff joints Warmness on joints Crackling sound when waking Diabetic Menopause
	Prostate gland Leg Deformation Hypertension Depression Good gait Asthma Weariness
	Anxiety

4.5 Results of Model Formulation

The next process, after feature selection used in identifying the most relevant variables among the 36 identified variables of KOA is model formulation, using the aforementioned supervised machine learning algorithms available in the Weka software. The 10-fold cross validation technique was used in validating the performance of the developed prognostic model for the risk of KOA using the test samples randomly selected from the historical test used for training the model. For each supervised machine learning algorithm used in formulating the prognostic model for the risk of KOA classification, 4 prognostic models were developed using the variables identified by each feature selection methods applied on the original dataset.

4.5.1 Result of Naïve Bayes Classifier and Screenshot

The results of the simulation of the prognostic model using Naïve Bayes classifier, in relative to the datasets of the originally identified 36 attributes and the relevant features selected using feature selection techniques are presented in the confusion matrices and model formulation screenshot shown in Figure 4.3 and Figure 4.4.

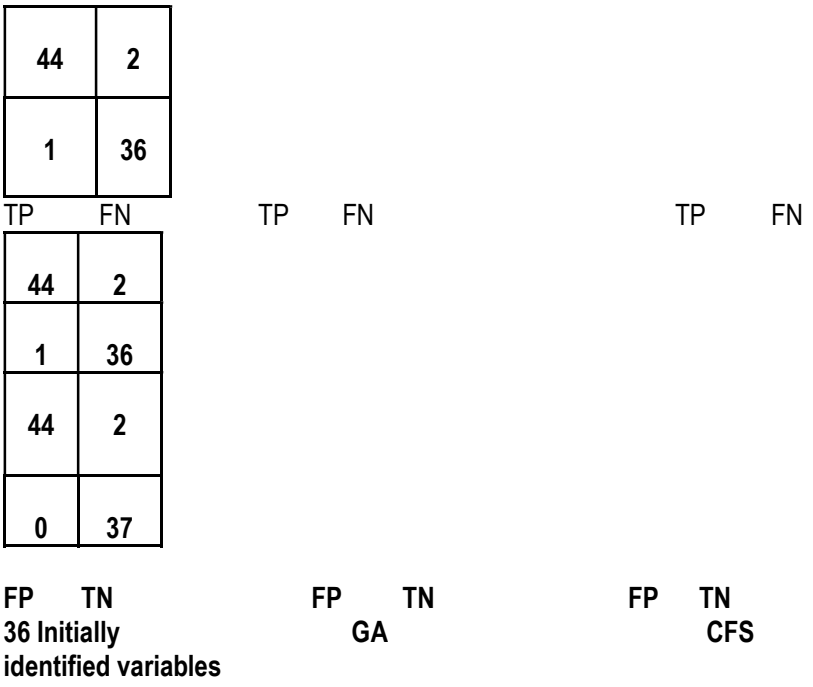

Figure 4.3: Performance of NB Classifiers Using Initial 36 Attributes, GA and CFS

TP = correctly classified Yes **FP = incorrectly classified No**
TN = correctly classified No **FN= incorrectly classified Yes**
Actual Cases = No= TN+TP **Yes = TP+FN**
Predicted Cases = No = FN+TN **Yes=TP+FP**

The confusion matrices shows the results of the NB classifier of the risk of knee osteoarthritis using 36 initially identified variables, 12 relevant variable by GA, and 7 relevant variables by CFS from left to right respectively. The correct classifications made by NB using 36 initially identified variables, GA, and CFS relevant variables selected was 81, 80 and 80 respectively while the misclassifications were 2, 3 and 3 owing for accuracies of 97.59%, 96.39% and 96.39 % respectively. Out of the actual 46 Yes cases, NB predicted 44, 44 and 44 correctly and out of the actual 37 No cases, NB correctly classified all 37 cases, 36 and 36 cases respectively.

The above analysis signifies that 46 out of the 83 KOA dataset collected were actual Yes (that is, patient affected with KOA). 44 out of actual Yes cases were correctly classified by \NB with no feature selection techniques applied. Applying G.A and CFS. 44 were also correctly classified respectively. Out of the 37 No cases of KOA (patient not affected with KOA), NB correctly classified all 37 No cases using all initially identified variables. Applying G.A and CFS 36 cases were correctly classified.

4.5.2 Result of C4.5 DT classifier and Screenshot

The results of the simulation of the prognostic model using C4.5DT classifier with relative to the datasets of the originally identified 36 attributes and the relevant features selected using feature selection techniques are presented in the confusion matrices and model formulation screenshot shown in Figure 4.5 and Figure 4.6.

The confusion matrices shows the results of the C4.5DT classifier of the risk of knee osteoarthritis using 36 initially identified variables, 12 relevant variable by GA, and 7 relevant variables by CFS from left to right respectively.

TP	FN
41	**5**
8	**29**
42	**4**
8	**29**
42	**4**
8	**29**

TP FN TP FN

FP TN FP TN FP TN

36 Initially identified variables **GA** **CFS**

Figure 4.5: Performance of C4.5DT Classifiers Using Initial 36 Attributes, GA and CFS

TP = correctly classified Yes FP = incorrectly classified No
TN = correctly classified No FN= incorrectly classified Yes
Actual Cases = No= TN+TP Yes = TP+FN
Predicted Cases = No = FN+TN Yes=TP+FP

The correct classifications made by C4.5 DT using 36 initially identified variables, GA, and CFS relevant variables selected was 71, 70 and 71 while the misclassifications were 12, 13 and 12 owing for accuracies of 85.54%, 84.3% and 85.54% respectively. Out of the actual 46 Yes cases of KOA, C4.5 DT, predicted 42, 41 and 42 correctly and out of the actual 37 No cases of KOA, C4.5 DT correctly classified 29, 29 and 29 cases respectively.

4.5.3 Result of MLP classifier and Screenshot

The results of the simulation of the prognostic model using MLP classifier relatively to the datasets of the originally identified 36 attributes and the relevant features selected using feature selection techniques are presented in the confusion matrices and model formulation screenshot shown in Figure 4.7 and Figure 4.8. The confusion matrices shows the results of the MLP classifier of the risk of knee osteoarthritis using 36 initially identified variables, 12 relevant variable by GA, and 7 relevant variables by CFS from left to right respectively.

The correct classifications made by MLP using 36 initially identified variables, GA, and CFS relevant variables selected was 81, 79 and 75 while the misclassifications were 2, 4 and 8 owing for accuracies of 97.59 %, 95.18% and 90.36% respectively. Out of the actual 46 Yes cases of KOA, MLP predicted 44, 44 and 43 correctly and out of the actual 37 No cases of KOA, MLP correctly classified all 37, 35 and 32 cases respectively.

4.5.4 Result of SVM classifier and Screenshot

The results of the prognostic model using SVM classifier with respect to the datasets of the originally identified 36 attributes and the relevant features selected are presented in the confusion matrices and model formulation screenshot shown in Figure 4.9 and Figure 4.10. The confusion matrices shows the results of the SVM

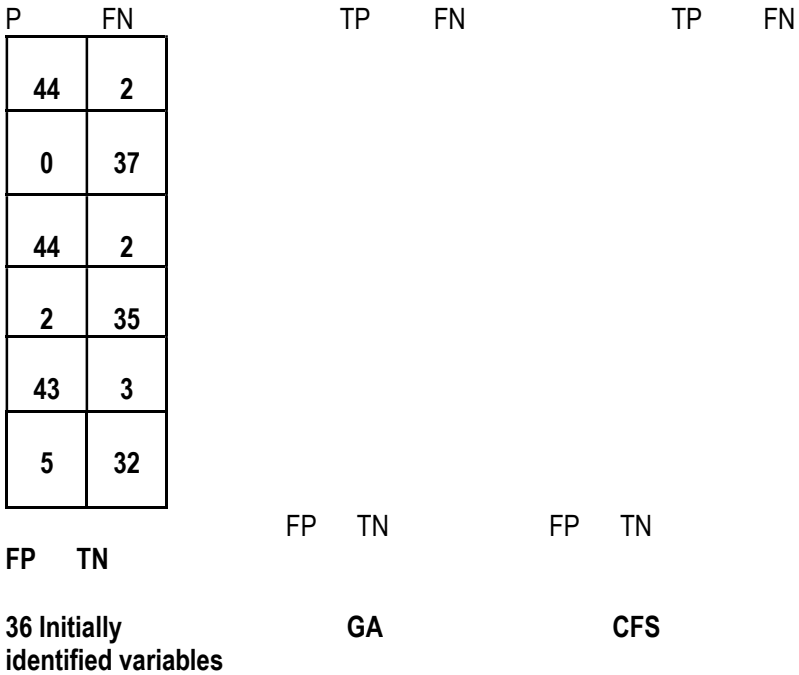

Figure 4.7: Performance of MLP Classifiers Using Initial 36 Attributes, GA and CFS

TP = correctly classified Yes **FP = incorrectly classified No**
TN = correctly classified No **FN= incorrectly classified Yes**
Actual Cases = No= TN+TP **Yes = TP+FN**
Predicted Cases = No = FN+TN **Yes=TP+FP**

TP	FN	TP	FN	TP	FN
44	2				
3	34				
43	3				
4	33				
44	3				
3	34				

FP	TN	FP	TN	FP	TN
36 Initially identified variables		GA		CFS	

Figure 4.9: Performance of SVM Classifiers Using Initial 36 Attributes, GA and CFS

TP = correctly classified Yes **FP = incorrectly classified No**
TN = correctly classified No **FN= incorrectly classified Yes**
Actual Cases = No= TN+TP **Yes = TP+FN**
Predicted Cases = No = FN+TN **Yes=TP+FP**

The classifier of the risk of knee osteoarthritis using 36 initially identified variables, 12 relevant variable by GA, and 7 relevant variables by CFS from left to right respectively. The correct classifications made by SVM using 36 initially identified variables, GA, and CFS relevant variables selected was 78, 76 and 77 while the misclassifications were 5, 7 and 6 owing for accuracies of 93.98%, 91.57% and 92.77% respectively. Out of the actual 46 Yes cases of KOA, SVM, predicted 44, 43 and 44 correctly and out of the actual 37 No cases of KOA, SVM correctly classified 34, 33 and 34 cases respectively.

4.6 Discussion of Prognostic Model

For each prediction model developed using the combination of feature selection and supervised machine learning algorithms; the confusion matrices were constructed from the value of correct (true positive and true negative values) and incorrect classifications (false positive and false negative values) made by each prediction model developed for risk of KOA. The positive class for each prediction model was identified by the Yes cases while the negative class for each prediction model was identified by the No cases.

The true positive and true negative values were used to evaluate the accuracy of each prognostic model showing how much of the total number of cases that was correctly classified by the classifiers – efficiency of the model. Additional metrics were estimated including the true positive rate which measures the ability of the model to correctly classify the Yes cases, true negative rate which measures the ability of the model to correctly classify the No cases, false positive rate which measures the incorrectly classified negative cases.

The NB classifier showed a relatively high level of performance irrespective of the type of feature selection method used in extracting the relevant variables; with 97.59%, 93.39% and 96.39% accuracy although, MLP also recorded high performance accuracy with no feature selected. In all type feature selection techniques, NB also predicted the highest Yes cases of 44 respectively and 37, 36 and 36 for the No cases. The NB classifier outperformed the DT, MLP and SVM algorithms while using all 36 variables and using 12 relevant variables selected by GA and 7 variables consistency-based feature selection algorithm.

Table 4.5: Summary of Evaluation Performance Metrics for the Models with no feature selection

Feature Selection Technique	Machine Learning Algorithm	Accuracy	TP rate		FP rate		Precision	
			Yes	No	Yes	No	Yes	No
No Feature Selection (36 variables considered)	NB	97.59	0.9565	0.4568	0.0000	0.0435	1.0000	0.9487
	DT	85.54	0.9130	0.4085	0.2162	0.0870	0.8400	0.8788
	MLP	97.59	0.9565	0.4568	0.0000	0.0435	1.0000	0.9487
	SVM	93.98	0.9565	0.4359	0.0811	0.0435	0.9362	0.9444

Table 4.6: Summary of Evaluation Performance Metrics for the Models using Genetic Algorithm

Feature Selection Technique	Machine Learning Algorithm	Accuracy	TP rate		FP rate		Precision	
			Yes	No	Yes	No	Yes	No
Genetic Algorithm (12 variables considered)	NB	96.39	0.957	0.450	0.027	0.044	0.978	0.947
	DT	84.34	0.891	0.414	0.216	0.108	0.837	0.853
	MLP	95.18	0.957	0.443	0.054	0.044	0.957	0.946
	SVM	91.57	0.935	0.434	0.108	0.065	0.915	0.917

Table 4.7: Summary of Evaluation Performance Metrics for the Models using Consistency based Feature

Feature Selection Technique	Machine Learning Algorithm	Accuracy	TP rate		FP rate		Precision	
			Yes	No	Yes	No	Yes	No
CFS (7 variables considered)	NB	96.39	0.957	0.450	0.027	0.044	0.978	0.947
	DT	85.54	0.913	0.409	0.216	0.087	0.840	0.879
	MLP	90.36	0.935	0.427	0.135	0.065	0.896	0.914
	SVM	92.77	0.935	0.442	0.081	0.065	0.935	0.919

The C4.5DT classifier had the least performance among the four classifier considered for this study. Using all the variables identified, that is, the 36 attributes, DT had 85.54% accuracy. Using GA, DT had 84.34% and 85.54% using CFS. Out of the 46 actual yes cases, C4.5DT correctly classified 42, 41 and 42 respectively using all features selected, G.A and CFS. Also, out of the 37 actual no cases, C4.5DT correctly classified 29 respectively for all features selection techniques used and when no feature was selected. A total of 71, 70, 71, cases were correctly classified and 12, 13, 12 were misclassified using No feature selection, G.A and CFS respectively.

MLP classifier performance is reasonably good too with respect to the feature selection techniques used, but not as good as the NB classifier. The accuracy percentage of MLP were 97.59, 95.18 and 90.36 respectively considering all 36 attributes, G.A and CFS. Out of the 46 actual yes cases, MLP correctly classified 44 for no feature selection techniques and GA, and 43 for CFS. A total of 81, 79, 75, cases were correctly classified and 2, 4, 8 were misclassified using no feature selection, G.A and CFS respectively.

SVM classifier accuracy were 97.59, 95.18 and 90.36 respectively considering all 36 attributes, G.A and CFS. Out of the 46 actual yes cases, SVM correctly classified 44 for no feature selection techniques, 43 for GA, and 44 for CFS. A total of 78, 76, 77, cases were correctly classified and 5, 7, 6 were misclassified using no feature selection, GA and CFS respectively.

Considering the performance per feature selection techniques, MLP and NB had the best and same performance followed by SVM and the least is DT. In GA feature selection techniques, NB maintain the best performance, followed by MLP, then SVM, the least classifier was DT.

Table 4.8: Evaluation of the Performance of Model Validation

Feature Selection Technique	Machine Learning Algorithm	Correct	Incorrect	Accuracy %	TP rate		FP rate		Precision	
					Yes	No	Yes	No	Yes	No
No Feature Selection	NB	81	2	97.59	0.957	1.000	0.000	0.043	1.000	0.949
	DT	71	12	85.54	0.913	0.784	0.216	0.087	0.840	0.879
	MLP	81	2	97.59	0.957	1.000	0.000	0.043	1.000	0.949
	SVM	78	5	93.98	0.957	0.919	0.081	0.043	0.936	0.944
Genetic Algorithm	NB	80	3	96.39	0.957	0.450	0.027	0.044	0.978	0.947
	DT	70	13	84.34	0.891	0.414	0.216	0.108	0.837	0.853
	MLP	79	4	95.18	0.957	0.443	0.054	0.044	0.957	0.946
	SVM	76	7	91.57	0.935	0.434	0.108	0.065	0.915	0.917
Consistency-Based Feature Selection	NB	80	3	96.39	0.957	0.450	0.027	0.044	0.978	0.947
	DT	71	12	85.54	0.913	0.409	0.216	0.087	0.840	0.879
	MLP	75	8	90.36	0.935	0.427	0.135	0.065	0.896	0.914
	SVM	77	6	92.77	0.935	0.442	0.081	0.065	0.935	0.919

Using CFS, NB still maintained the best performance followed by SVM, then MLP and lastly DT . produce optimum result. The results showed that the highest Precision values for Yes cases were recorded for using the NB classifier with and without feature selection (FS) with values of 1 while the highest Precision for No cases was recorded for using NB classifier without feature selection with values of 0.949. Therefore, the highest Precision for Yes and No cases was achieved for using NB or MLP classifier without feature selection. The model using no feature selection (36 variables) had the best performance for NB and MLP although, there was no much difference when feature selection was applied with NB topping the list for GA and CFS with 96.4% and 96.4% accuracy .

However feature selection was able to select relevant attributes that could be helpful when monitoring the risk of KOA in a patient. NB outperformed DT, MLP and SVM when considering the overall performance, followed by MLP, then SVM and lastly DT. The following are the seven (7) relevant attributes listed by CFS in order of their importance; Age, LGA of residence, pain (while climbing staircase), weight, pain (when joints are pressed), visible swelling on joints and gait and the twelve (12) relevant attributes selected by GA in order of their importance are Age, LGA of residence, pain (while climbing staircase), weight, pain (while walking), pain (when

joints are pressed), visible swelling on joints, warmness on joints, feeling weary or nervous, menopause, leg deformation and gait. A close observation on the attributes selected by the two feature selections reveals that all the features selected by CFS are among the ones selected by GA, also the order of importance for the first to four attributes are in the same order for G.A and CFS.

The distribution of the results derived from the evaluation of the performance by class (risk and no risk) alongside the number of correct classifications made with the respective accuracy were also display in Figure 4.11. The bar chart shows the distribution of the number of correct classifications (blue bars) alongside the respective accuracy (red bars measured as a percentage). The results of the study showed that the highest values were recorded for using the NB classifier using feature selection (FS) with values 96.4% however using the initially identified features, the NB and MLP had the best accuracies with values of 97.6% by each. The bar chart in Figure 4.12 shows the distribution of the TP rate for Yes and No (blue and red bars respectively) alongside the Precision for Yes and No (green and purple bars respectively). The results showed that the highest TP rate values for Yes cases were recorded for using the NB classifier with and without feature selection (FS) with values of 0.957 while the highest TP rate for No cases was recorded for using NB classifier without feature selection with values of 1. Therefore, the highest TP rates for Yes and No cases were achieved for using NB or MLP classifier without feature selection.

5. CONCLUSION

In conclusion, having identifying the variables relevant to the risk of KOA; formulated models using four SML classifiers, simulated the model using weka simulation software as well as validated the performance of the model with 10 fold cross validation technique, it could be inferred from this study that all 36 identified attributes are relevant for predicting the risk of KOA. Also, variables such as gait, menopause, sport, warmness of joint and leg deformation were absent in some of the existing models meanwhile, these variables are equally relevant for developing a KOA prognostic model. All the feature selections identified age as the most important variable for KOA. This condition affect more male than female in this study and affect adults more. The prognostic model developed using the datasets showed good results although there was more likelihood to get better performance if dataset is increased. Report from this study has estimated trends in patients' outcome and serves as a tool for monitoring the risk of having KOA.

REFERENCES

1. Akinpelu, O.A., Alonge, T.O., Adekanla, B.A. and Odole, A.C. (2009). Prevalence and Pattern of Symptomatic Knee Osteoarthritis in Nigeria: A Community Based Study. The Internet Journal of Allied Health Sciences and Practice 7(3):1504–80.
2. Ali G., Saeed S., Mohammad N., and Sayed S. H. (2015). The Applications of Genetic Algorithms in Medicine. Oman Med J. 30(6): 406–416.
3. Arthritis Foundation (2017). What is Osteoarthritis? Available from http:// www.arthritis.org/about-arthritis/types/osteoarthritis [Accessed on 23/6/2017]
4. Black, J.E., Terry, A.L. and Lizotte, D.J (2017). FRAMR-EMR: Framework for Prognostic Predictive Model Development Using Electronic Medical Record Data with a Case Study in Osteoarthritis Risk. Available from https://arxiv.org/ftp/arxiv/papers/1705/1705.09563.pdf [Access on 3/4/2018]
5. Chu, C.R, Williams, A.A, Coyle, C.H. and Bowers, M.E. (2012). Early Diagnosis to Enable Early Treatment of Pre-Osteoarthritis. Arthritis Research and Therapy. 14: 212-213

6. Deo, R.C. (2015). Machine Learning in Machine. Available from https://www.ncbi.nlm.nih.gov/pubmed [Accessed on 3/2/18]
7. Gabay, O and Clouse, K.A. (2016). Epigenetics of Cartilage Diseases. Joint Bone Spine Newsletter. 3(5):6
8. Gardiner, B.S., Woodhouse, F.G., Besier, T.F., Grodzinsky, A.J., Lloyd, D.G., Zhang, L. and Smith, D.W. (2016). Predicting Knee Osteoarthritis. Annals of biomedical Engineering. 44: 222–233
9. Kerkar. P. (2017). Osteoarthritis or Wear And Tear Arthritis: Types, Causes, Symptoms, Treatment-Surgery Available From Https://Www.Epainassist.Com/Arthritis/Osteoarthritis-Or-Wear-And-Tear-Arthritis [Accessed on 3/12/2017]
10. Kerkhof, H. J. M. Bierma-Zeinstra, S. M. A., Arden, N. K., Metrustry S. Castano-Betancourt, M., Hart, D J., Hofman, A., Rivadeneira, F., Oei, E H., Spector, T.D., Uitterlinden, A. G., Janssens, A. C. J. W. Valdes, A. Meurs, M. J. Van, B. J (2013). Prediction Model for Knee Osteoarthritis Incidence, Including Clinical, Genetic and Biochemical Risk Factors. Annals of Rheumatic Disease -2013-203620
11. Kontzias, A. (2017). Osteoarthritis (OA) (Degenerative Joint Disease; Osteoarthrosis; Hypertrophic Osteoarthritis) Available from http://www.msdmanuals.com/professional/musculoskeletal-and-connective-tissue-disorders/joint-disorders/osteoarthritis-oa [Accessed on 29/11/2017]
12. Kraus, V. B., Blanco, F. J., Englund, M., Karsda, M. A., and Lohmander, L. S. (2015). Call for standardized definitions of osteoarthritis and risk stratification for clinical trials and clinical use. Osteoarthritis. Cartilages. 2015:03.036.
13. Kumar, K.V., Shyamalaa, K. and Nareshc, D. (2017). Prediction Model on Knee Osteoarthritis. International Science Press ISSN: 0974–5572.10(23)
14. Lazzarini, N., Runhaar, J., Bay-Jensen, A., Thudium, C., Bierma-Zeinstra, S., Henrotin, Y., Bacardit, J., (2017). A Machine Learning Approach for the Identification of New Biomarkers for Knee Osteoarthritis Development in Overweight and Obese Women. Osteoarthritis cartilage. 2(12):2014-2021
15. Litwic, A., Edwards, M.H., Dennison, E.M .and Cooper, C. (2013) Epidemiology and Burden of Osteoarthritis. British Medical Bulletin. 105: 185-199
16. Lohmander S.L. (2000) What Can We Do About Osteoarthritis? Arthritis Research and Therapy. 2 (2): 95-100.
17. Longton, W., Kira, R., Shinaman, R., Celis, E. and Coughlan, J. (2016).Osteoarthritis Pain Management Available from http://www.painmedicineconsultants.com /conditions-osteoarthritis.htm [Accessed on 23/06/2017]
18. Maiese, K. (2016). Picking a Bone with WISP1 (CCCN4): New Strategies against Degenerative Joint Disease. Journal of Translation Science. 1(3): 83–85.
19. Mirer, F.E. and Stellman, J.M. (2008). Occupational Safety and Health Protections International Encyclopaedia of Public Health : 658–668
20. Mitchell, M. (1996). An Introduction to Genetic Algorithms. MIT Press. Cambridge, massachusets, England.5th Ed. 112-120
21. Mitchell, T. (1997).Machine Learning, McGraw-Hill Science/Engineering, Portland. 2nd Ed. 219-225.
22. Murray, C.J., Vos, T., Lozano, R., Naghavi, M., Flaxman, A.D. and Michaud, C. (2013). Disability-Adjusted Life Years for 291 Diseases and Injuries in 21 Regions, 1990–2010: A Systematic Analysis for the Global Burden of Disease Study 2010. Lancet, 380:2197–2223.
23. Persson, P.V. and Rietz, H. (2017). Predicting and Analyzing Osteoarthritis Patient Outcomes with Machine Learning. Master's Thesis. Lund University. ISSN 1650-2884.

24. Riviere, C. (2017).Why Knee Osteoarthritis? Available from http://www.charlesriviere.co.uk/knee/patient-education/why-knee-osteoarthritis/ [Accessed on 8/12/17]
25. Staal, A.V. (1999). Predictive Models in Medicine: Some Methods for Construction and Adaptation. Unpublished PhD thesis. Submitted to the Norwegian University of Science and Technology .Available from, http//www.citeseerx. ist.psu.edu. [Accessed on 4/4/2018]
26. Symmons, D., Mathers, C., and Pfleger, B. (2006). Global Burden of Osteoarthritis in the Year 2000. World Health Organization (report 2006). Available from http://www.who.int/healthinfo/statistics/bod_osteoarthritis.pdf [Accessed on 4/12/17]
27. Symmons, D., Mathers, C., and Pfleger, B. (2015). Global Burden of Rheumatoid Arthritis in the Year 2000. WHO Report 2006. Available from www.who.int/healthinfo/statistics/bod_rheumatoidarthritis.pdf. [Accessed on 3/9/17]
28. United States Bone and Joint Initiative (USBJI) (2015) What Is The Impact Of Burden Of Musculoskeletal Disorders And Why Is The U.S. Bone And Joint Initiative Important? Bone and Joint initiative USA Available from http://www.boneandjointburden.org/. [Accessed on 2/2/18].
29. Waijee, A., Mukherjee, A. and Singal, A. (2013). Comparison of Modern Imputation Methods for Missing Laboratory Data in Medicine. British Medical Journal. 3 (8): 1–7.
30. Woolf, A. D. and Pfleger, B. (2003). Burden of Major Musculoskeletal Conditions Bulletin of the World Health Organization 81:646-656.
31. World Health Organization (2012). “Chronic Rheumatic Conditions.” Chronic Diseases and Health Promotion. Available from http://Www.Who.Int/Chp/Topics/ Rheumatic/En/ 2/8/17 [Accessed on 3/8/17]
32. Yarnell, J. and O’Reilly, D. (2013). Epidemiology and Disease Prevention: A Global Approach OUP Oxford University Press, England. (2nd Ed). 54-60
33. Zhang, S., Zhang, C. and Yang, Q. (2002). Data Preparation for Data Mining. Applied artificial Intelligence. 17: 375 – 381.
34. Zhang, W., Mcwilliams, D.F., Ingham, S.L., Doherty, S.A., Muthuri, S., Muir, K.R., and Doherty, M. (2011). Nottingham Knee Osteoarthritis Risk Prediction Models. Annals of the Rheumatic Diseases. 70(9):1599-604.
35. Zhang, Y. and Jordan, J.M. (2010). Epidemiology of Osteoarthritis. Clinics in geriatric medicine. 26: 355-369.