Volume 9, Issue 3, March – 2024 International Journal of Innovative Science and Research Technology
ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR2040

An Early Predictive Model for the Onset of Knees Osteoarthritis in Nigeria

Oladejo, Rachel Adefunke[1]; Engr. Oyedeji Ayo Isaac[2]; Engr. Oluleye Gabriel[3]; Engr. Akinrogunde Oluwadare Olatunde[4]; Adenle Bamidele. J[5]
Computer Science Department[1]; Computer Engineering Department[2]; Electrical Electronics Department[3,4]; Computer Software Engineering Department[5]
Ogun State Institte of Technology, Igbesa,Ogun State, Nigeria[1,2,3,4]
Dots ICT Institute of Technology[5]

Abstract:- The primary risk factors for patients with Knee Osteoarthritis (KOA) were determined in this study, and a predictive model was developed using the data found. In order to comprehend the body of information regarding musculoskeletal-related diseases, a thorough study of relevant literature was conducted. One ailment that falls within the musculoskeletal category is knee osteoarthritis, and the risk factors were extracted and confirmed by medical professionals. clinical data encompassing characteristics tracked during KOA patients' treatment were gathered from Ile-Ife, Osun State, Nigeria at the OAU Teaching Hospital Complex (OAUTHC), , as well as from a few other people Utilizing questionnaires, . For this investigation, the entire dataset comprising data on 83 patients was used. WEKA software was used to compare four supervised machine learning techniques so as to create the model. The accuracy of the was 97.59% when examining the 36 originally identified attributes without selecting any featue. The outcomes additionally demonstrated The minimal amount of variables pertinent to the osteoarthritis condition of the knee. Subsequent findings demonstrated the relevance of each feature found in order to create a prognosis model for knee osteoarthritis that is both effective and efficient. Age is the most important factor for KOA, according to the study's findings, and all 36 characteristics were found to be useful in forecasting the likelihood of Knee Osteoarthritis..

Keywords:- *Prognostic Model, Supervised Machine Learning, Knee Osteoarthritis.*

I. INTRODUCTION

The most prevalent musculoskeletal ailment is osteoarthritis (OA), often known as wear and tear disease or degenerative joint issue. This condition depreciates affected cartilage over time. ([1];[2]). In a typical joint, cartilage covers each bone's end. It is a tough, rubbery material. It acts as a cushion between the bones and enables easy bone gliding. But when OA progresses, cartilage or pieces of bone begin to break, resulting in pain, swelling, and the inability to move the affected joint [3]. This either floats around the joint or grows into spurs. As osteoarthritis progresses, cartilage gets deteriorated, causing the bones to scrape against one another. Joint injury and increased discomfort result from this [4]. Despite spending billions on research, there are currently no medications that have been proved to alter the biological course of OA, and there are only a few treatments that have been shown to be effective. Currently, OA is incurable and no medication has been provided for repair of harms that Osteoarthritis causes [6]. Due to the fact that osteoarthritis cannot be reversed, its frequency rises steadily with age. The development and progression of osteoarthritis can be influenced by old age, obesity, heredity, gender, bone density, trauma, and lack of exercise [7]. OA is physically, psychologically, and socioeconomically taxing. Immobility, that is, decreasing movement as well as day-to-day life tasks, may be linked to it. A few of the psychological repercussions are loneliness, diminished self-worth, and distress. Due to the high prevalence of OA in the population, it has a large economic impact [8].

Despite the abundance of publications on the increasing frequency of musculoskeletal diseases, there are few and underreported reports from Africa. Rheumatoid arthritis's prevalence in Africa was estimated in 2006 using research from South Africa, Nigeria, and Liberia [9]. According to literature and global trends, this revealed a high male-to-female ratio [9]. Similar to this, only one South African study was utilized to estimate the prevalence of osteoarthritis in Africa, underscoring the dearth of information on that continent [10]. 19.6% of persons in Nigeria who are 40 or older and have symptoms of knee osteoarthritis [11] have this condition. Despite extensive and expensive research that has cost exorbitantly, medications has not been provided that can alter the course, and just a little therapies have been shown to relieve it.

In spite KOA begin a major causal of immobility, no model exists at the moment that takes into account the necessary amount of variables to forecast it. To solve this issue, an algorithm (model) that will help to assess risk of knee osteoarthritis using necessary risk factors must be developed. By the time this model is put into use, it might be integrated into existing health information systems, which would have an impact on the analysis of KOA clinical data in real-time. This would also benefit patients and other stakeholders as it will make essential decisions

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR2040

and resource allocation easier to regions deemed important to address prevalence..

II. REVIEW OF RELATED WORKS

Several studies on the categorization, diagnosis, and risk prediction of osteoarthritis have been published. But the list below includes a handful of the already-published works.

A machine learning-based prediction model for incidence radiographic osteoarthritis (OA) of the knee over an 8-year period was built by Joseph, McCulloch, Nevitt, and Link [13] using four variables and only three models that were compared. In order to categorize the presence and 8-year incidence of knee pain using MRI, Lee, Liu, Majumdar, and Pedoia [14] used data-driven feature learning. Then, using sagittal intermediate-weighted 2D turbo spin-echo fat suppression MRI, a 3D DenseNet was trained to identify the existence of discomfort. Next, functional principal component analysis (FPCA) was used to predict the temporal patterns of pain incidence among non-symptomatic knees at baseline. To locate groups of pain trajectories, Bayesian Gaussian mixture models were fitted to the FPCA scores. The 3D DenseNet learned cluster membership from MRI alone and MRI paired with clinical characteristics. Cluster membership was used as labels. Although other samples were looked at, the research is focused on medical imaging. By integrating logistic regression analysis, Receiver Operating Characteristic (ROC), Integrated Discrimination Improvement (IDI), and Kaplan-Meier curves with just seven (7) characteristics, Lourido et al. [15] established a model to predict the prospective development of radiographic KOA (rKOA). Using 3D gait analysis and 3T magnetic resonance imaging (MRI), Atkinson et al. [16] investigated the relationship between change in surrogate measures of knee load and knee effusion-synovitis in patients with medial compartment knee OA receiving high tibial osteotomy (HTO). According to the study's findings, a decrease in medial knee load is positively correlated with a decrease in knee inflammation following HTO, pointing to the possibility of mechano-inflammatory in people with knee OA Machine learning was used by Persson and Rietz [17] to predict and analyze osteoarthritis patient outcomes. 15 variables were taken into account. Five methods (Logistic Regression, Random Forest, Adaptive Boosting, Gradient Boosting, and Multi-Layer Perceptron) were compared without the use of feature selection strategies to locate relevant features. Gradient Boosting model produced the best results. Fewer factors were included, and certain crucial variables, such menopause, family history, and gait, were not taken into account.

Models for predicting the risk of Nottingham knee osteoarthritis were studied in 2011 [18]. Knee osteoarthritis (OA) risk prediction model was created. In addition, it was estimated how much risk might be reduced by changing potential risk factors. Only a logistic regression model, which took into account nine (9) variables, was employed. There were fewer variables included, and certain crucial variables like Menopause, Leg deformity, and Gait were not taken into account. A knee osteoarthritis prediction model was created using machine learning by [19]. Using a comparison of the two models, Logistic regression (LR) and Naive Bayes (NB), it was determined how likely it was for a patient to acquire OA. The prediction model that fits the data the best is provided by logistic regression. This research has value.

A case study on the risk of osteoarthritis and studies on a framework for creating prognostic, predictive models utilizing data from electronic medical records were both completed in 2017. As a result, a Framework for a Prognostic Model was developed, which provides detailed instructions for developing a prognostic, predictive model using EMR data based on variables such as BMI, gender, age, osteoporosis and history of knee injuries, and LR predicted osteoarthritis, but only a small number of factors were considered, and important factors like occupation, sport, and leg deformity were not employed [20].

Most contemporary models have been developed using few risk variables (important variables such as family history, sport, leg deformation, climbing stairs, menopause and good gait were omitted not *considered*).

A large number of these models are from other countries, and as a result, the results may be influenced by factors such as varied environments, diets, climatic conditions, occupations, and access to healthcare. Because of this, our study stands out among all other predictive models for KOA.

III. METHOD

One of the methods used to build the prediction model to forecast the risk of KOA among persons in Nigeria involved a thorough evaluation of relevant studies to discover and explore the various risks factors for knee osteoarthritis. For validating the risk variables found in the literature purpose, five (5) physiotherapists were questioned. Additionally, pertinent information from hospital medical records and questionnaires was gathered. This information contained the risk factors required for tracking knee osteoarthritis. The prediction of Knee Osteoarthritis was done using a Genetic Algorithm that was developed based on supervised machine learning algorithms. The model was simulated using the Waikato Environment for Knowledge Analysis (WEKA) program. The performance of the prognostic model was confirmed by calculating false-positive rate, true positive rate, accuracy and precision using the historical data acquired.

A. Identification of Data and Data Gathering

In the study, eighty-three (83) data points were observed and evaluated, including some sick and healthy individual data;

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR2040

Table 1: Identified Clinical and Demographic Factors

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 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

B. Feature Selection Using Genetic Algorithm

The relevant variables were chosen from the first sets detected using Genetic Tool (G.A), a meta-heuristic computational intelligence algorithm. According to equation (1), employing the I initially detected variables may result in performance that is equal to or superior than using the selected r attributes such that $X_r \subset X_i$.

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

Therefore, GA chosen in this research had been used in changing dataset with i attribute and n records to dataset with r attributes and n records such that $r < i, X^{n\times i}$ is a data matrix having *n* records with *i* initially identified variables while $X^{n\times r}$ is a data matrix having *n* records with *r* reduced attributes and $i, n \; and \; r \in \mathbb{Z}^+$. according to equation (2)

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

In this study, the most pertinent variables were selected from the initial list of variables using GA, a population-based search heuristic algorithm that simulates the natural evolution process.

GA used a process of natural selection mixed with mutation and crossover techniques to create a new population by employing one population of chromosomes (referred to as the solution candidates). A fitness function or objective function was used to assess the chromosomes' fitness.

This study shows that a gene value of "1" signifies that a certain attribute indexed by position "1" is chosen. contrary, a gene value of "0" denotes the selection of the

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR2040

attribute indexed at position "0". Equation (3) shows that the value of the fitness of chromosomes selected can be evaluated using a fitness function. The values show the KNN classification error as well as N_f as number of attributes that were chosen.

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

The remaining individuals in the current population are used to produce the rest of the next generation through crossover and mutation, after the elite individuals are moved to the next generation. 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

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

On the other hand, mutation was carried out by flipping the bit strings in accordance with the likelihood of mutation given to the employed GA. The GA method left behind the best chromosome, allowing positions of the most important characteristics that were chosen by the GA to be identified by the index of the bit string for which there was a value of "1". Table 3 lists the settings that were applied to the GA that was suggested for feature selection in this investigation. The datasets were reduced to one that had the pertinent qualities related to KOA as a result of applying GA to the dataset containing the first identified variables.

C. Model Formulation

The prognostic model needed to assess the risk of KOA was created using supervised classification techniques using the identified variables related to that risk.

Since the study's goal is to determine if a risk of KOA exists or not, the necessary task is a classification issue that identifies the outcomes of the values of the set of attributes provided as an output for defining each KOA risk.

The created supervised machine learning algorithms must translate the values of the input attribute set X to the target class set Y, which consists of Yes or No cases, as shown by equation (5).

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

Table 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

D. Stochastic-based Supervised Machine Learning Algorithms

A predictive model for the likelihood of developing knee osteoarthritis was developed using the stochastic-based Machine Learning (ML) algorithms Naive Bayes and C4.5 decision tree classifiers. The next paragraphs provide a presentation of the algorithms.

➢ *Naïve Bayes' (NB) classifier for the risk KOA*

One of the most popular techniques for supervised learning is the naïve Bayes' Classifier, a probabilistic model built on the Bayes' theorem of conditional probability. Based on probability theory, it offers an effective method for managing any number of attributes or classes. Practical learning techniques and prior knowledge of the observed data are provided by Bayesian classification. Let X_{ij} be a dataset sample with records (or occurrences) of I different qualities related to the risk of developing knee osteoarthritis along with their relative risk of developing knee osteo arthritis. *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*. Nave Bayes' classification requires the determination of the following in order to classify the risk of knee osteoarthritis given the values of the risk factor of the jth record:

- $P(H_k|X_{ij})$ – Posteriori probability: the likelihood that the hypothesis H_k will hold given the sample of observed data 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*;
- $P(|X_{ij}|H_k)$ is the probability of observing the sample's attribute, X_i given that the hypothesis holds in the training data X_{ij}.

Hence, the hypothesis' posteriori probability Equation (6) illustrates the Bayes' Theorem for each class. Equation (7) then calculates the chance of developing knee osteoarthritis based on the results for each class in equation (6)

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

$$Risk\ of\ Osteoarthritis max.\left[P(H_1|X_{ij}), P(H_2|X_{ij})\right] \quad (7)$$

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR2040

➢ *C4.5Decision Trees Classifier for the Risk of KOA*

The decision tree is a supervised machine learning algorithm that uses a divide-and-rule strategy to grow a recursive hierarchical tree. This tree can be thought of as a collection of If-Then statements or rules that combine the attributes in order to predict the likelihood of developing knee osteoarthritis. To do this, the training dataset was divided into subsets based on an attribute value test for each input variable, and the tree used recursive partitioning to repeat the procedure on each subset to learn the pattern in the dataset. When the subset at a node contains every member of the target class or when splitting no longer provides value, the recursion is finished.

The C4.5 decision tree's halting criteria were the Gain ratio, which at each node generation calculates the information gain of each characteristic specified in equation (8) and divides it by the split value in accordance with equation (10).

Therefore, the attribute X_i with the greatest value of the gain, the 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:

$$H(X_i) = -\sum_{t \epsilon T} \frac{|t, X_i|}{|X_{ij}|} \cdot \log_2 \frac{|t, X_i|}{|X_{ij}|} \quad (9)$$

$$Split(T) = -\sum_{t \epsilon T} \frac{|t|}{|X_{ij}|} \cdot \log_2 \frac{|t|}{|X_{ij}|} \quad (10)$$

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 were reached.

E. Perceptron-based Supervised Machine Learning Algorithm

For this study, Support Vector Machine and Multilayer Perceptron (MLP) were considered for formulating a prognostic model for the risk of knee osteoarthritis.

➢ *Support Vector Machines (SVM) for the risk of KOA*

During model formulation, SVM attempted to minimize the cost of classification by optimizing the distance between hyper-planes.

Given that the higher the margin, the lower the generalization error of the SVM classifier, the hyperplane $< \boldsymbol{w}, \boldsymbol{x} > + b = 0,$ with the longest distance $\frac{2}{||w||}$ to the neighboring data points of either class at opposing ends accomplished a satisfactory separation.

hyperplane created was defined as $< \boldsymbol{w}, \boldsymbol{x} > + b = 0$ where $\boldsymbol{w} \epsilon \mathbb{R}^p$ and $b \epsilon \mathbb{R}$ while $< \boldsymbol{w}, \boldsymbol{x} > + b = -1$ and $< \boldsymbol{w}, \boldsymbol{x} > + b = 1$ are the margins required for the separation *w* of support vectors *x* within the *n* variables. Hence, the decision function in Eq. (13) was used to propagate the output of Eq. (12) using a sigmoid function with an interval of { -1, 1}. Eq. (12) was utilized to define a linearly separable function. The SVM seeks to maximize the separation of the hyper-planes in equation (13), subject to the decision function specified in equation (14).

$$Risk_i = f(X_i) = (< W, X_i > + b) > 0, \quad \forall i \epsilon [1, n] \quad (12)$$

$$f_d(x_i) = sign(Risk_i) = (< W, X_i > + b) > 0, \quad \forall i \epsilon [1, n] \quad (13)$$

$$maximize \ \frac{1}{2} ||\boldsymbol{w}||^2 \quad (14)$$

➢ *Artificial Neural Network – Multi-layer perceptron (MLP) for the risk of Knee Osteoarthritis*

- Phase 1 - Propagation: Each propagation entails the subsequent steps:
- Forward propagation of the input from the training pattern through each node j in the neural network to produce the output activations of the propagation;

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

- The neural network's training pattern target is used to create deltas δ_j for all of the output and hidden neurons as the output activations are propagated backward through it.
- Phase 2 - Weight update: As a result of each weight synapses, we have the following:
- To determine the gradient of the weight, multiply its output delta and input activation.

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

- Deduct from the weight a ratio (percentage α) of the gradient.

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

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR2040

➢ *Performance Evaluation Metrics*

- Plotting the classification results on a confusion matrix was necessary to assess how well the supervised machine learning algorithms performed in classifying the risk of developing knee osteoarthritis. In order to assess the performance of the prediction model using performance evaluation metrics, the true positive/negative and false positive/negative values recorded from the confusion matrix were used. The following is a description of the metrics' definitions and expressions:
- **True Positive (TP) rates (sensitivity/recall)**-The percentage of positively diagnosed cases that were accurately identified.

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

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

- False Positive (FP) rates (1-specificity/false alarms) – the percentage of negative cases that are mistakenly labeled as positives (1-specificity/false alarms).

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

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

- Precision-the percentage of correctly anticipated positive/negative cases.

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

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

- Accuracy – the proportion of the total correct predictions.

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

The performance of the prognostic model for the classification of risk of knee osteoarthritis was evaluated by validation using a historical dataset gathered based on the information supplied in the questionnaire, using the performance criteria indicated above. The TP rate and precision are both between [0, 1], whereas the accuracy is between [0, 100]% and the FP rate is between [0, 1]. The better the model, the nearer to 100% accuracy the accuracy is. The greater the TP rate and precision values, and the worse the FP rate values, the closer they are to 1 and 0 respectively. As a result, a successful model has high TP/Precision rates and low FP rates when it is evaluated

IV. RESULTS AND DISCUSSION

A. *Results of Data Identification and Collection*

Thirty-six (36) variables were chosen for this study, and 83 questionnaires were completed by the participants and respondents. In tables 4,5 and 6, descriptive statistical frequency distribution tables were used to analyze these replies in order to see how the knee osteoarthritis risk was distributed across the study participants. Based on the data gathered, it was found that 46 (55.4%) records contained patients at risk for developing knee osteoarthritis, while 37 (44.6%) records contained patients at low risk.

Table 4: Results of the Demographic Variables Description

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)

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR2040

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)
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 5: Results of the Summary Statistics for the Numerical 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 6: Results of the Associated Variables' Identification and Description

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)

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR2040

	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)
Stiff Joints	Yes	35 (42.2)
	No	45 (54.2)
	Missing	3 (3.6)
Warmness on joints	Yes	15 (18.1)
	No	65 (78.3)
	Missing	3 (3.6)
Crackling sound when walking	Yes	9 (10.8)
	No	72 (86.7)
	Missing	2 (2.4)
Diabetic	Yes	6 (7.2)
	No	76 (91.6)
	Missing	1 (1.2)

Table 6(b): Continuation of the Results of the Associated Variables' Identification and Description

Menopause	Yes	23 (27.8)
	No	17 (20.5)
	NA	43 (51.9)
Prostate gland	Yes	4 (4.8)
	No	38 (45.8)
	NA	40 (48.2)
	Missing	1 (1.2)
Leg deformation	Yes	10 (12.0)
	No	70 (84.3)
	Not sure	3 (3.6)
Hypertensive	Yes	26 (31.3)
	No	55 (66.3)
	Missing	2 (2.4)
Depression	Yes	4 (4.8)
	No	74 (89.2)
	Missing	5 (6.0)
Good gait	Yes	60 (72.3)
	No	23 (27.7)

B. Results of Feature Selection Process

Two feature selection methods (Genetic Algorithm and Consistency base feature Selection) were applied to the initially identified variables to remove the variables with a high correlation and add the variables that have a high correlation with the target variables.

➢ *Feature Selection Using Genetic Algorithm*

The genetic algorithm was used to extract the most important features from the initial 36 attributes that had been found, and the original dataset was fed to the algorithm. This was done by subjecting the genetic algorithm to the process of searching through the entire attribute space of possible subsets (2^{36}) of selected attributes. Each subset of attributes that were collected was represented using a chromosome bit string such that if an attribute is selected, a value of 1 was provided on the index else 0. Following the selection of the initial population, the fitness of each of the selected attributes was evaluated using the KNN classifier, which was embedded into the GA process. After that, the process for which the GA was used to extract the most relevant features (some attributes) was identified.

➢ *Consistency based Feature Selection (CFS)*

Another feature selection technique was used in addition to the Genetic Algorithm to choose the variables in order to find pertinent variables. It was done using a subset evaluator that ordered the qualities according to importance, like the consistency-based feature selection technique.

For the purpose of measuring the efficacy of the prognostic models, the set of attributes identified by each FS algorithm in addition to the initially detected attributes were passed to the supervised machine learning algorithms.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR2040

The model's performance was then developed using the pertinent features that had been chosen, and the 36 variables were contrasted to see which performed the best.

C. Results of Model Formulation

Using the supervised machine learning algorithms included in the Weka software, the model formulation process comes next. Using test samples randomly chosen from the historical test used to train the model, the 10-fold cross-validation technique was utilized to validate the performance of the proposed prognostic model for the risk of KOA. The variables discovered by each feature selection technique applied to the original dataset were utilized to develop 4 prognostic models for each supervised machine learning methodology.

➢ *Result of Naïve Bayes Classifier*

The Naive Bayes (NB) classifier of the risk of developing knee osteoarthritis used 36 initially identified factors, 12 relevant variables by GA, and 7 relevant variables by CFS, respectively. The findings are displayed in the confusion matrices. Using the 36 initially identified factors, the GA, and the CFS relevant variables, the correct classifications made by NB were 81, 80, and 80, respectively, whereas the misclassifications were 2, 3, and 3 due to accuracy errors of 97.59%, 96.39%, and 96.39%, respectively. In the actual 46 Yes cases, NB properly predicted 44, 44, and 44; in the actual37 No cases, NB correctly identified all 37 cases36 and 36 cases.

➢ *Result of C4.5 DT classifier and Screenshot*

Using 36 initially identified factors, 12 relevant variables by GA, and 7 relevant variables by CFS, respectively, the confusion matrices display the outcomes of the C4.5DT classifier of the risk of developing knee osteoarthritis.

Table 7: 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
		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 walking Diabetic Menopause
		Prostate gland Leg Deformation Hypertension Depression Good gait Asthma Weariness
		Anxiety

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR2040

The C4.5 DT correctly classified 71, 70, and 71 out of the 36 initially identified factors, the GA, and the relevant CFS variables, while misclassifying 12, 13, and 12 with accuracy errors of 85.54%, 84.3%, and 85.54%, respectively. Out of the actual 46 KOA yes cases, C4.5 DT accurately predicted 42, 41, and 42; similarly, out of the actual 37 KOA no cases, C4.5 DT properly identified 29, 29, and 29 cases respectively.

- *Result of MLP Classifier*

Using 36 initially identified variables, 12 relevant variables by GA, and 7 relevant variables by CFS, respectively, the confusion matrices display the outcomes of the MLP classifier of the risk of developing knee osteoarthritis.

The 36 initially identified factors, GA, and CFS pertinent variables were used to classify the data by MLP. The accurate classifications were 81, 79, and 75; however, the misclassifications were 2, 4, and 8 due to accuracy errors of 97.59%, 95.18%, and 90.36%, respectively. Of of the actual 37 No cases of KOA, MLP properly identified all 37, 35, and 32 cases, while out of the actual 46 Yes cases of KOA, MLP accurately predicted 44, 44, and 43 cases.

➢ *Result of SVM Classifier*

The correct classifications made by SVM using the 36 initially identified variables, GA, and CFS relevant variables selected were 78, 76, and 77, while the misclassifications were 5, 7, and 6 owing to accuracies of 93.98%, 91.57%, and 92.77%, respectively. SVM correctly identified 44, 43, and 44 of the actual 46 KOA yes cases, and 34, 33, and 34 of the actual 37 KOA no cases.

D. Validation and Evaluation performance of the model formulation

The confusion matrices were constructed from the values of the correct classifications (true positive and true negative values) and incorrect classifications (false positive and false negative values) made by each prognostic model developed for risk of KOA for each prognostic model developed using the combination of feature selection and supervised machine learning algorithms. For each prognosis model, the actual yes instances indicated the positive class, and for each prediction model, the actual no cases identified the negative class.

As previously noted, a 10-fold cross-validation technique was used in validating the performance of the created prognostic model for the risk of KOA using test samples randomly picked from the historical test used to train the model.

Table 8: 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

The true positive and true negative values were used to assess each prognostic model's efficacy by displaying the proportion of the total number of instances that were correctly classified by the classifiers. Other metrics were calculated, such as the true positive rate, which gauges how well the model categorizes cases that actually answer yes, and the false positive rate, which gauges cases that are mistakenly labeled as positive. Precision measures the proportion of accurately anticipated cases, either positive or negative, and accuracy, which assess the fraction of the total correctly predicted.

Table 8 displays the final model's outcomes. This table displays the results of the evaluation of the model's performance using data gathered based on the original risk factors along with pertinent features chosen using GA and consistency-based feature selection algorithms. The models were developed using four algorithm to determine the approach that produces the best result.

V. DISCUSSION

The confusion matrices were built constructed from the values of the correct classifications (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 for each prediction model developed using the combination of feature selection and supervised machine learning algorithms.

The Yes cases identified the positive class for each prediction model, and the No cases revealed the negative class for each prediction model.

The true positive and true negative values were used to evaluate each prognostic model's efficacy by indicating how much of the total number of cases were correctly identified by the classifiers. Other metrics were computed, such as the true positive rate, which evaluate the model's accuracy in classifying yes cases, and the true negative rate, which evaluate the model's accuracy in classifying no cases.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR2040

Table 9: Evaluation Performance Metrics for Models using Genetic Algorithm Summarized

	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 10: Evaluation Performance Metrics for Models Using Consistency-Based Feature Summary

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	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

Table 11: 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 (36 variables considered)	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

Correctly, and false-positive rate, which measures the incorrectly classified negative cases. Regardless of the feature selection technique used to extract the pertinent variables, the NB classifier demonstrated a rather high degree of performance, achieving accuracy rates of 97.59%, 93.39%, and 96.39%. MLP did, however, also record great performance accuracy without any features turned on. In all feature selection methods, NB also predicted the greatest Yes cases—44 in each case—and the most No cases—37, 36, and 36. Using all 36 variables, 12 relevant variables chosen by GA, and 7 consistency-based feature selection algorithms, the NB classifier beat the DT, MLP, and SVM algorithms. Tables 9, 10, and 11 display this.

Out of the four classifiers taken into consideration for this investigation, the C4.5DT classifier had the worst performance. DT had an accuracy of 85.54% when using all of the detected factors, or the 36 attributes. DT achieved 84.34% when using GA and 85.54% when utilizing CFS. Using all features chosen, G.A. and CFS, C4.5DT accurately categorized 42, 41, and 42 of the 46 actual yes cases. A feature was not chosen in any of the 37 actual no situations. In total, 71 instances were correctly classified utilizing No feature selection, G.A., and CFS, while only 12, 13, and 12 cases were incorrectly identified using these methods.

The performance of the MLP classifier in terms of the feature selection algorithms used is likewise respectably good, though not as good as the NB classifier. The accuracy% of MLP was 97.59, 95.18, and 90.36 when all 36 characteristics, including G.A. and CFS, were taken into account. Out of the 46 actual affirmative cases, MLP properly recognized 44 for no feature selection techniques, GA, and CFS 43. With no feature selection, G.A., and CFS, respectively, 81, 79, and 75 cases were wrongly classified, whereas 2, 4, and 8 cases were correctly classified in the remaining cases.

The accuracy of the SVM classifier with 36 features, G.A., and CFS was 97.59, 95.18, and 90.36, respectively. SVM correctly classified 43 of the 46 genuine yes cases for GA and 44 of the 46 genuine yes cases for CFS for no feature selection procedures. 78, 76, and 77 instances were correctly classified using no feature selection, GA, and CFS, whereas 5, 7, and 6 cases were wrongly classified.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR2040

MLP and NB had the best and same-performing feature selection algorithms, followed by SVM, and DT had the worst results. MLP, SVM, and DT as the least effective classifier follow NB as the top performers in GA feature selection approaches. While dealing with CFS, NB continued to perform best, followed in that order by SVM, MLP, and DT. The results showed that for Yes cases, the NB classifier with and without feature selection (FS) produced the highest Precision scores. In contrast, the NB classifier without feature selection produced the highest Precision for No cases, with values of 0.949.

Thus, the NB or MLP classifier without feature selection produced the greatest Precision for Yes and No cases. The model with 36 variables and no feature selection performed the best for NB and MLP. When feature selection was used, there was no difference, with NB coming out on top for GA and CFS with 96.4% 96.4% accuracy.

Nonetheless, feature selection was able to choose pertinent qualities that could be useful for assessing a patient's risk for KOA. When comparing overall performance, NB performed better than DT, MLP, and SVM, and DT came in bottom.

VI. CONCLUSION

It can be inferred from this study that all 36 identified attributes are significant after identifying the variables that are crucial for estimating the risk of KOA, developing models using four SML classifiers, simulating the model using weka simulation software, and validating the model's performance using the 10-fold cross-validation technique. Variables like gait, menopause, sport, the warmth of joints, and leg deformation were absent from a few of the earlier models. These elements are equally crucial for developing a prognostic model for KOA.

In all feature options, age was consistently regarded as the most important aspect by KOA. This disease affects more males than females and more adults in this study. The prognostic model created using the datasets yielded promising results, although performance was more likely to advance with more datasets. The report from this study provides an estimated trend in patient outcomes and a tool for monitoring the risk of developing KOA.

REFERENCES

[1]. Kontzias, "Osteoarthritis (OA) (Degenerative Joint Disease; Osteoarthrosis; Hypertrophic Osteoarthritis)," [Online]. Available: http://www.msdmanuals.com/professional/musculoskeletal-and-connective-tissue-disorders/joint-disorders/osteoarthritis-oa . [Accessed 29 11 2017].

[2]. Kerkar, "Osteoarthritis or Wear And Tear Arthritis: Types, Causes, Symptoms, Treatment- Surgery," 2017. [Online]. Available: Https://Www.Epainassist.Com/Arthritis/Osteoarthritis-Or-Wear-And-Tear-Arthritis . [Accessed 3 12 2017].

[3]. Arthritis Foundation , "What is Osteoarthritis?," 2017. [Online]. Available: http:// www.arthritis.org/about-arthritis/types/osteoarthritis . [Accessed 23 6 2017].

[4]. C. Riviere, "Why Knee Osteoarthritis?," 2017. [Online]. Available: http://www.charlesriviere.co.uk/knee/patient-education/why-knee-osteoarthritis/. [Accessed 8 12 2017].

[5]. W. Longton, R. Kira, R. Shinaman, E. Celis and J. Coughlan, "Osteoarthritis Pain Management," Painmedicineconsultants, 2016. [Online]. Available: http://www.painmedicineconsultants.com /conditions-osteoarthritis.htm. [Accessed 23 6 2017].

[6]. O. Gabay and K. Clouse, "Epigenetics of Cartilage Diseases," Joint Bone Spine Newsletter., London, 2016.

[7]. Litwic, M. Edwards, E. Dennison and C. .. Cooper, "Epidemiology and Burden of Osteoarthritis.," *British Medical Bulletin*, no. 105:, pp. 185-199, 2013.

[8]. Symmons, C. Mathers and B. and Pfleger, "Global Burden of Rheumatoid Arthritis in the Year 2000. WHO Report 2006," WHO, britain, 2015.

[9]. D. Symmons, C. Mathers and B. Pfleger, "Global Burden of Osteoarthritis in the Year 2000.," World Health Organization (report 2006), 2006.

[10]. O. Akinpelu, T. Alonge, B. Adekanla and A. Odole, "Prevalence and Pattern of Symptomatic Knee Osteoarthritis in Nigeria: A Community Based Study.," *The Internet Journal of Allied Health Sciences and Practice*, vol. 7, no. 3, p. 1504–80., (2009.

[11]. B. Gardiner, F. Woodhouse, T. Besier, A. Grodzinsky, D. Lloyd, L. Zhang and D. Smith, "Predicting Knee Osteoarthritis.," *Annals of Biomedical Engineering*, vol. 44, p. 222–233, 2016.

[12]. P. Persson and H. Rietz, Predicting and Analyzing Osteoarthritis Patient Outcomes with Machine Learning., Master's Thesis: Lund University., 2017.

[13]. W. Zhang, D. Mcwilliams, S. Ingham, S. Doherty, S. Muthuri, K. Muir and M. Doherty, "Nottingham Knee Osteoarthritis Risk Prediction Models.," *Annals of the Rheumatic Diseases*, vol. 70, no. 9, pp. 599-604, 2011.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR2040

[14]. K. Kumar, K. Shyamalaa and D. Nareshc, "Prediction Model on Knee Osteoarthritis.," *International Science Press,* no. ISSN: 0974–5572.10(23), 2017.
[15]. J. Black, A. Terry and D. Lizotte, "FRAMR-EMR: Framework for Prognostic Predictive Model Development Using Electronic Medical Record Data with a Case Study in Osteoarthritis Risk.," 2017.
[16]. J. Lee, F. Liu, MajumdarS. and V. Pedoia, "An ensemble clinical and MR-image deep learning model predicts 8-year knee pain trajectory," *osteoarthritis initiative,Osteoarthritis Imaging,,* vol. 100003, no. ISSN 2772-6541, p. 1, 2021.
[17]. Joseph, McCullochC.E., M. Nevitt, T. Link and J. John, "Machine learning to predict incident radiographic knee osteoarthritis over 8 Years using combined MR imaging features, demographics, and clinical factors:," *Osteoarthritis Initiative, Osteoarthritis and Cartilage,* vol. 30, no. 2, pp. 270-279, 2021.
[18]. Atkinson, T. Birmingham, C. Primeau, J. Schulz, C. Appleton, P. S.L. and J. Giffin, "Association between changes in knee load and effusion-synovitis: evidence of mechano-inflammation in knee osteoarthritis using high tibial osteotoosteotomy as a model.," *Osteoarthritis and cartilage*, Vols. 22-29, no. 2021, p. 29, 2020.
[19]. L. Lourido, V. Balboa-Barreiro, C. Ruiz-Romero, I. Rego-Pérez, M. Camacho-Encina, R. Paz-González, V. Calamia, N. Oreiro, P. Nilsson and F. Blanco, "A clinical model including protein biomarkers predicts radiographic knee osteoarthritis: a prospective," *Osteoarthritis Initiative,Osteoarthritis and Cartilage,,* vol. 29, no. 8, pp. 1147-115, 2021.