IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-11, Nov 2019 E-ISSN: 2347-2693

A Model to Detect Heart Disease using Machine Learning Algorithm

O.E. Taylor[1*], P. S. Ezekiel[2], F.B. Deedam-Okuchaba[3]

[1,2,3]Dept. of Computer Science, Rivers State University, Port Harcourt, Nigeria

[*]*Corresponding Author: tayonate@yahoo.com, Tel.: +2348034448978*

DOI: https://doi.org/10.26438/ijcse/v7i11.15 | Available online at: www.ijcseonline.org

Accepted: 09/Nov/2019, Published: 30/Nov/2019

***Abstract*—** Heart disease also refers to conditions that involve narrowed or blocked blood vessels that can lead to a heart attack, chest pain (angina) or stroke. This paper presents a model for detecting heart disease using machine learning algorithm. The methodology adopted in this research is Agile Methodology, which follows planning, requirements analysis, designing, coding, testing and documentation in parallel during the stage of production process. In this paper a Heart Dataset was trained using four different machine learning algorithms (K-Nearest Neighbours Classifier, Support Vector Classifier, Decision Tree Classifier and Random Forest Classifier). The algorithm with the best accurate result was used in making predictions. This model was deployed to the web using flask (a python framework), it takes 13 inputs from the user in order to make prediction. The model is implemented using python programming language and flask (a web base framework). This paper uses a Decision Tree Classifier Algorithm and the results obtained from the prediction shows an accuracy of about 98.83%, which is really encouraging.

***Keywords*—** Heart Disease, Machine Learning, K-Nearest Neighbors, Support Vector machine, Decision Tree, Random Forest

I. INTRODUCTION

Heart disease is one of the leading causes of death and hospitalization in both genders in nearly all countries in Africa, thus substantially representing a public health risk or burden. Given the pressing need to implement comprehensive strategies to address this growing epidemic, surveillance remains the primary tool to evaluate the burden of the disease, to assess the growing trend, plan preventive actions at both population as well as individual levels and to estimate efficacy of prevention. The most frequent Heart Diseases are those of atherosclerotic origin, mainly Ischemic Heart Disease (IHD) and stroke. Heart Disease clinically manifests itself in middle life and also at an older age, after many years of exposure to unhealthy lifestyles which includes (unhealthy diet, physical inactivity, and smoking habit) and risk factors (high blood pressure, high cholesterolemia, diabetes, obesity etcetera). Although the prevalence of it is very high, its occurrence is largely preventable, making it a priority for public health and sustainability. Epidemiological studies have demonstrated that cardiovascular risk is 'reversible', that means that by lowering the level of risk factors it is possible to reduce the number and severity of events, or delay the event occurrence. Even though the clinical onset is mainly acute, Heart Diseases often evolve gradually, causing substantial loss of quality of life, disability, and lifelong dependence on health services and medications. This will eventually lead to premature death as well as adverse outcomes in elderly people, including cognitive impairment, dementia and decreased physical performance. The societal costs of Heart Disease are substantial and include not only those directly related to health care and social services, but also those linked to illness benefits and retirement, impact on families and caregivers, and loss of years of productive life.

Heart Disease has been identified as one of the largest causes of death even in developed countries [2]. The application of Machine learning based heart disease detection and prediction system were discussed in several research findings. The application of artificial intelligence in disease detection systems especially the cardiac disease system detection which improves the performance of other existing widely used models like models provided by American College of Cardiology/American Heart Association (ACC/AHA) models in CVD detection and prediction [1]. This paper presents a model for detecting heart disease using machine learning algorithm.

II. RELATED WORK

The survey on heart disease prediction system based on data mining was carried out by [3]. The paper highlighted the use of data mining in discovering trends in patient data through pattern generation. This technique enhanced and improved their health strategy. The algorithms they presented were with a specific end goal to anticipate the coronary illness which included some constraint. They produced an affiliation

guideline on a genuine informational index with the patients' history in regards to coronary illness to yield high exactness rate. The proposed calculation they carried out handles the issue of vast number of principles and appropriate approval of guidelines backings as well. Kernel F-score Feature Selection was introduced to perform determination as a pre-preparing venture in the characterization of therapeutic database. The proposed KFFS technique includes two stages; the first was to change the components of the medical datasets to bit space stage by methods for Linear or RBF capacities. The second was Utilizing F-score equation; the therapeutic datasets have been ascertained by changing piece capacities from non-directly detachable medical dataset to a straight distinguishable element space [3].

In the paper titled "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques" [4], data mining classification techniques, namely Decision Trees, Naive Bayes, and Neural Networks were analysed on Heart disease database. The performance of various techniques used in the research was compared, based on accuracy. The results accuracy of Neural Networks, Decision Trees, and Naive Bayes were 100%, 99.62%, and 90.74% respectively. Their analysis shows that out of these three classification models, Neural Networks predicts Heart disease with the highest accuracy. Finally, they added two more input attributes namely; obesity and smoking. Both attributes were used to get more accurate results due to their importance in detecting heart diseases.

In the paper "Intelligent Heart Disease Prediction System Using Data Mining Techniques", a prototype Intelligent Heart Disease Prediction System (IHDPS) as developed using data mining techniques, namely, Decision Trees, Naïve Bayes and Neural Network. Their results show that each technique has its unique strength in realizing the objectives of the defined mining goals. IHDPS can answer complex "what if" queries which traditional decision support systems cannot. Using medical profiles such as age, sex, blood pressure and blood sugar it can predict the likelihood of patients getting a heart disease. It enables significant knowledge, for example patterns, relationships between medical factors related to heart disease, to be established [5].
In the paper "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques"; the study carried out an investigation using a method termed ensemble classification, which is used for improving the accuracy of weak algorithms by combining multiple classifiers. Experiments with this tool were performed using a heart disease dataset. A comparative analytical approach was done to determine how the ensemble technique can be applied for improving prediction accuracy in heart diseases. The results of these study indicates that the ensemble techniques, such as bagging and boosting, are effective in improving the prediction accuracy of weak classifiers, and exhibit satisfactory performance in identifying the possible risk of having a heart disease. A maximum increase of 7% accuracy for weak classifiers was achieved with the help of the ensemble classification. The performance of the process was further enhanced with a feature selection implementation, and the results showed significant improvement in prediction accuracy [6].

III. METHODOLOGY

The methodology used here is the Rapid Application Development (RAD) and Prototype Design Specification (PDS). RAD methodology is designed to be flexible to changes and to accept new inputs, like features and functions, at every step of the development process

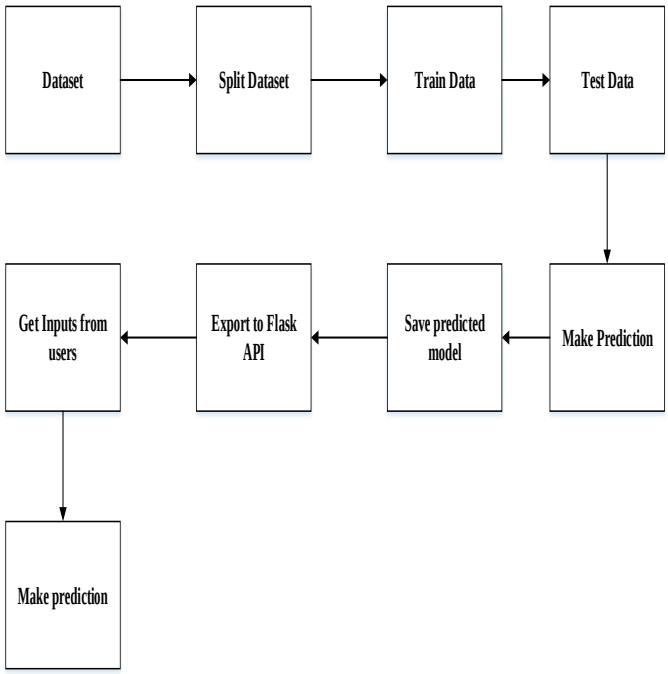

Figure 1: Architecture of the proposed system model

This system uses a dataset called heart dataset which was gotten from kaggle.com. This dataset contains 14 attributes of test results carried out on 1025 persons. The dataset was split into a train and a test sets. Here, this dataset was trained using four machine algorithms which namely: K-Nearest Neighbours Support Vector Classifier, Decision Tree Classifier and Random Forest Classifier. After training, testing and checking for accuracy of the four different algorithms used in training the model, the model with the highest percentage of accuracy was used to makes prediction. This trained model is saved and then loaded into a web. Flask which is an Application Programming Interface (AP1) loads this trained model and take inputs form from the users to fill in their test results. These inputs are being passed to the trained model to check and make prediction of a patient having a heart disease or not.

IV. RESULTS AND DISCUSSION

In this paper, a machine learning model was being trained in other to determine if a user has a heart disease or not. This machine model uses a dataset which have 13 test results conducted on different persons. This dataset was being cleaned and processed making sure that there are no null values present in the dataset. This dataset was split into x and y variables. Where the x variable contains the 13 attributes which are the different test results and the y variable contains the output. The x variable was being scaled using StandardScaler. The x variable and y variable were further divided into x_train, x_test, y_train and y_test. These x_train and y_train where being fitted or trained using four machine algorithms which are K-Nearest Neighbors, Support Vector Machine, Decision Tree and Random Forest. The four algorithms where used in checking the percentage of accurate results using different numbers of n values. For K Neighbors, the highest accurate result is 97.47% approximately when n =1, for Support Vector Machine, the highest accurate result is 98.83% when number of estimator =10, for Decision Tree, the highest accurate result is 98.83% when number of n =1, for Random Forest, the highest accurate result is 98.83% when number of estimator =10. After the testing of accuracy, we used Decision Tree Classifier which has one of the highest accurate results in making prediction. The Decision Tree model was being saved and loaded into the web using an Application Programming Interface called Flask. Using Flask, we created an HTML page containing 13 inputs of which the users will enter their different test result and pass the inputs to the model to detect if they have a Heart Disease or not. The result of the model will also be displayed on the web to the user.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal
0	52	1	0	125	212	0	1	168	0	1.0	2	2	3
1	53	1	0	140	203	1	0	155	1	3.1	0	0	3
2	70	1	0	145	174	0	1	125	1	2.6	0	0	3
3	61	1	0	148	203	0	1	161	0	0.0	2	1	3
4	62	0	0	138	294	1	1	106	0	1.9	1	3	2
5	58	0	0	100	248	0	0	122	0	1.0	1	0	2
6	58	1	0	114	318	0	2	140	0	4.4	0	3	1
7	55	1	0	160	289	0	0	145	1	0.8	1	1	3
8	46	1	0	120	249	0	0	144	0	0.8	2	0	3
9	54	1	0	122	286	0	0	116	1	3.2	1	2	2
10	71	0	0	112	149	0	1	125	0	1.6	1	0	2
11	43	0	0	132	341	1	0	136	1	3.0	1	0	3
12	34	0	1	118	210	0	1	192	0	0.7	2	0	2
13	51	1	0	140	298	0	1	122	1	4.2	1	3	3
14	52	1	0	128	204	1	1	156	1	1.0	1	0	0
15	34	0	1	118	210	0	1	192	0	0.7	2	0	2
16	51	0	2	140	308	0	0	142	0	1.5	2	1	2

Figure 2: Showing the training data of 1025 persons' test result which was been feed to four different machine algorithm for training.

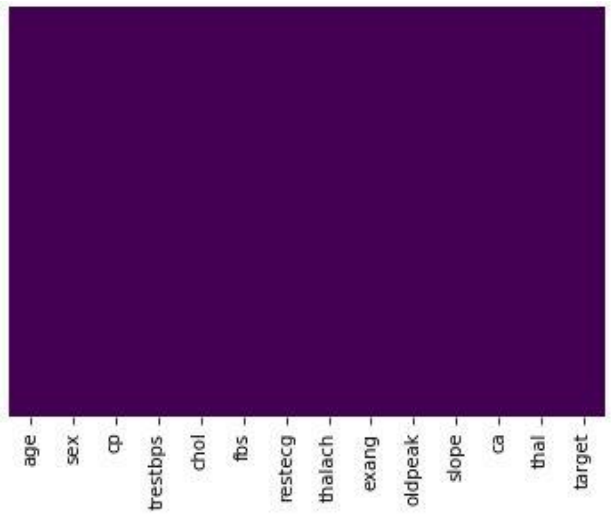

Figure 3: Showing that the dataset is totally cleaned indicating that there are no null values present

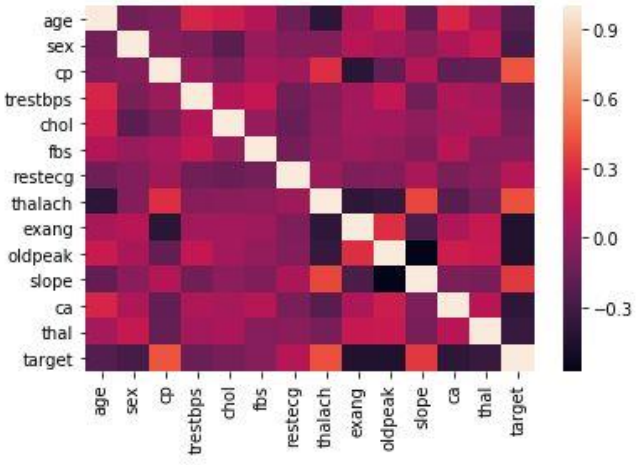

Figure 4: Plot showing a correlation matrix of the dataset

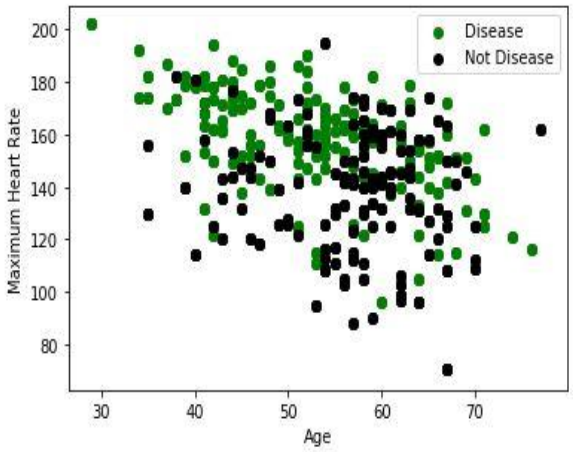

Figure 5: showing a scatter plot showing the maximum heart rate and the age of persons having a heart disease or not.

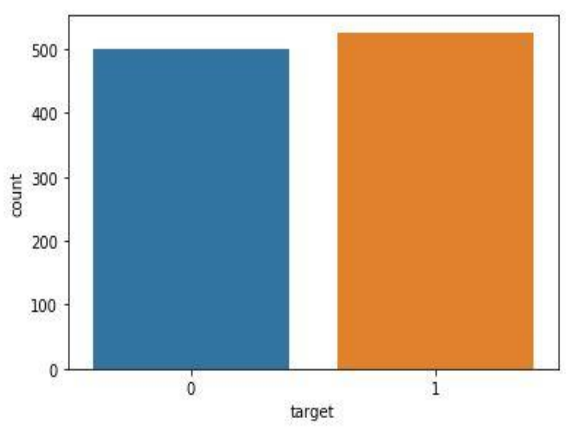

Figure 6: A histogram counting the number persons having a heart disease or not

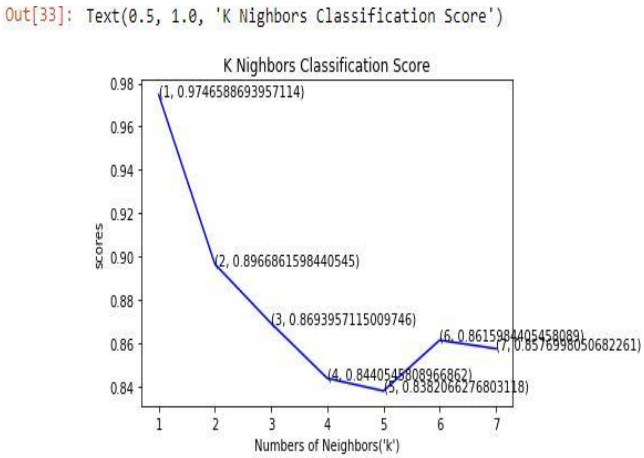

Figure 7: showing accurate scores of K Neighbors ranging from when n is 1 to when n is 7

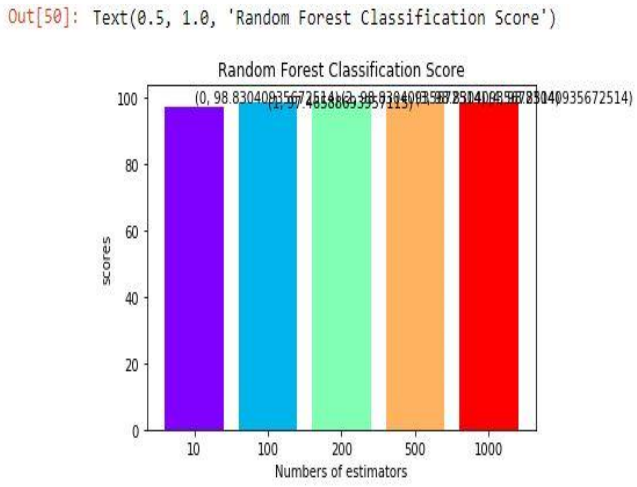

Figure 8: showing accurate scores of Support Vector Machine ranging from when n is 10, 100, 200, 500, 1000

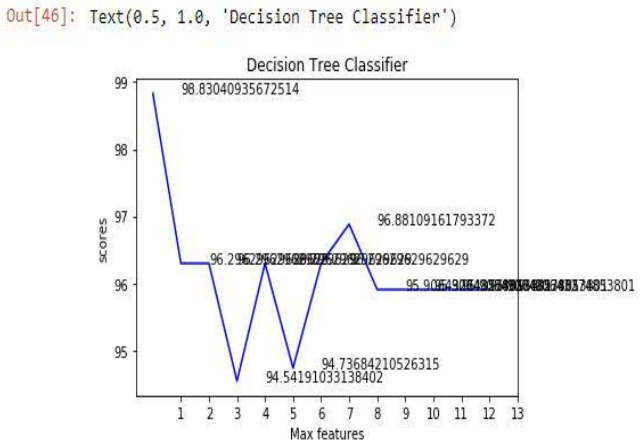

Figure 9: showing accurate scores of Decision Tree ranging from when n is 1 to when n is 13

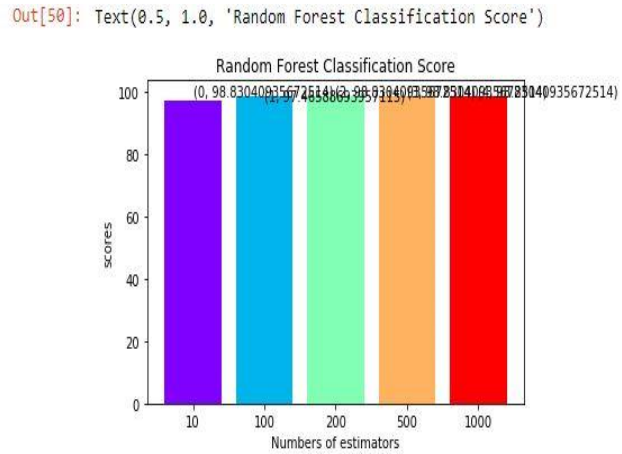

Figure 10: showing accurate scores of Random Forest ranging from when n is 10, 100, 200, 500, 1000

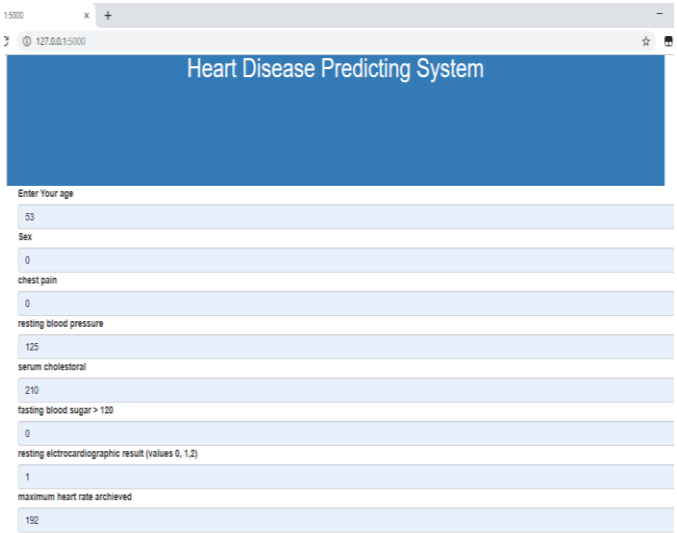

Figure 11: showing user input form to which patients will have to input some test results

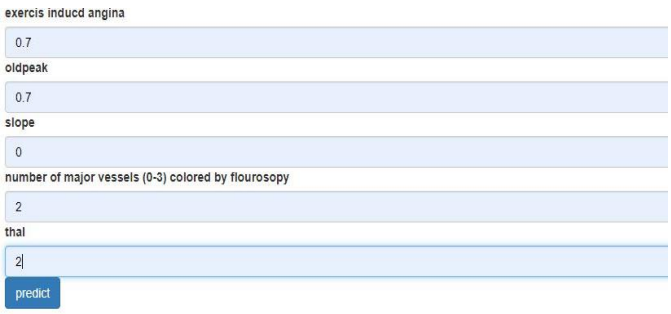

Figure 12: continuation of user input form to which users will have to input some test results

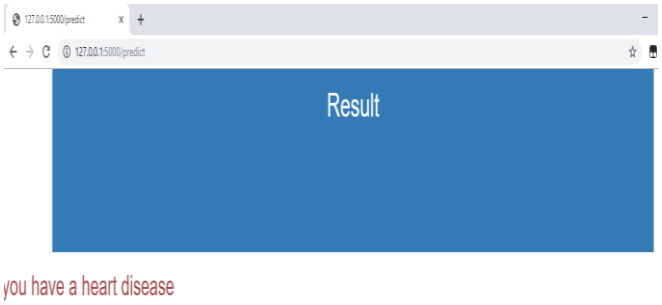

Figure 13: Results of the inputted test results showing that a user has a heart disease

V. CONCLUSION AND FUTURE SCOPE

This paper which bothered on proposing "A model to Detect Heart Disease Using machine learning Algorithm" was developed by using a machine learning approach. In this machine learning approach four algorithms were used to train and analyse the dataset which contains the test results of different patients and these algorithms were also tested for accuracy plotting a graph using matplotlib. After testing for accuracy, Decision Tree, Random Forest and Support Vector Machine have the highest accurate result which is about 98.83% approximately while K Nearest Neighbors have 97.4% approximately. Decision Tree model was also integrated in the web through an API called Flask, and it predicted good results when tested 5 times on the web without an error. This research can be extended to a real-time system using Deep Learning approach, where users can upload their test results as image.

REFERENCES

[1]. K. Vanisree, S. Jyothi, "*Decision Support System for Congenital Heart Disease Diagnosis based on Signs and Symptoms using Neural Networks",* International Journal of Computer Applications vol.**19**, issue.**6**, pp.6 – 12, **2011.**

[2]. S.F. Weng, J. Reps, J. Kai, J.M. Garibaldi, N. Qureshi, "*Can Machine-Learning Improve Cardiovascular Risk Prediction Using Routine Clinical Data*", vol.**1**, issue.**12**, pp. e0174944**, 2017.**

[3]. M. Thiyagaraj, G. Suseendran, "*Survey on heart disease prediction system based on data mining techniques*", Indian Journal of Innovations and Developments vol.**6** issue.**1**, pp.1-9, **2017**.

[4]. C.S. Dangare, S.S. Apte, "*Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques*", International Journal of Computer Applications vol.**47**, issue.**10**, pp. 44-48, **2012.**

[5]. S. Palaniappan, R. Awang, "Intelligent heart disease prediction system using data mining techniques", In *2008 IEEE/ACS international conference on computer systems and applications*, pp. **108-115**, **2008**.

[6]. C.B.C. Latha, S.C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques", *Informatics in Medicine, Unlocked 16*, pp.**100203**, **2019**.

Authors Profile

Mr. O. E. Taylor pursed his B.Sc degree in 2000 and M.Sc degree in 2004 all in Computer Science from the Rivers State University of Science and Technology and University of Ibadan, Nigeria respectively. He is currently a lecturer in the Department of Computer science, Rivers State University, Port Harcourt, Nigeeria. He is currently undergoing his Ph.D programme in Computer Science at the University of Port Harcourt, Nigeria. He is a member of the Computer Professionals (Registration Council) of Nigeria and Nigeria Computer Society. His research focuses on intelligent systems, smart space, context-aware sytems, machines learning algorithms and artificial intelligence. He has 14 years of teaching experience.

Mr. P E Ezekiel pursed his Bachelor of Science degree from Department of Computer Science, Rivers State University. He has published just a research paper in international journal of Computer Science and Mathematical Theory it's also available online. His main research work focuses on Machine Learning, Data Science, Deep Learning and Artificial Intelligence.

Mrs. F. B Deedam-Okuchaba pursed Her Bachelor of Science in Computer Science from the Rivers State University of Science and Technology (Now Rivers State Uniiversity) Nigeria in 2008 and Master of Science in Management Informaiton Systems from Coventry University, United Kingdom. She is currently working as Lecturer II in the Department of Computer Science, Rivers State University since 2015. She is a chartered member of Computer Professionals Registration Council of Nigeria (CPN) and Nigerian Women in Information Technology (NIIWIIT). She has two research publications and her research interest is in Electronic and Mobile Learning, Human Computer Interaction, Artificial Intelligence and Machine Learning. She has 4 years of teaching experience and 3 years of research expereince.

