Prediction Of Diabetes Using Machine Learning Algorithms

Obiabunmo Ugochukwu[1]
Department Of Electrical/Electronic and Computer Engineering
University of Uyo, Akwa Ibom State Nigeria
obiabunmou@hotmail.com

Constance Kalu[2]
Department Of Electrical/Electronic and Computer Engineering
University of Uyo, Akwa Ibom State Nigeria

Philip M. Asuquo[3]
Department Of Electrical/Electronic And Computer Engineering,
University of Uyo, Akwa Ibom State Nigeria
philipasuquo@uniuyo.edu.ng

***Abstract*— This study presents the prediction of diabetes using Pima Indians Diabetes Dataset which were pre-processed and then employed in training four different machine learning models, namely; Logistic Regression algorithm, Support Vector Machine (SVM) algorithm, Random Forest algorithm and K Nearest Neighbors (KNN) algorithm. The algorithms were trained in K-5 folds and their performance matrix which included average value, accuracy, precision, F1-score and recall were obtained. The result revealed that the Random Forest outperformed the other models with the highest accuracy of 81.25 percent, Precision of 79.57 percent, Recall of 77.98 percent and F1-Score of 78.73 percent. Hence, the Random Forest algorithm was recommeneded for for detecting patients with diabetes or patients with the likelihood of having diabetes. The ideas presented in this work has can assist the medical experts to predict the likelihood of having diabetes and to identify some hidden patterns in the factors that cause diabetes.**

Keywords— Support Vector Machine model, diabetes , Logistic Regression model, Pima Indians Diabetes Dataset, and Random Forest models.

1. Introduction

Nowadays, technology driven solutions are used to address proffer solutions to problems in virtually every human endeavor [1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22]. Also, applications of different diverse algorithms and approaches have facilitated technological solutions that cuts across data acquisition, modelling, prediction, analysis, forecasting, and visualization, among others [21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39, 40]. Accordingly, in this paper, the focus is on data driven model for prediction of diabetes. Natively, diabetes is one of the world's most serious health issues. There are four categories of diabetes, and they are: Type 1 diabetes, type 2 diabetes, gestational diabetes, and other kinds of diabetes [41,42,43,43]. Diabetes, in any form, raises the risk of long-term consequences. These usually appear after a number of years (10–20), but they may be the initial symptom in those who have not yet been diagnosed. The World Health Organization (WHO) established guidelines for diagnosing diabetes in pregnancy in 2006 [45]. According to 2017 statistics, almost 425 million individuals have diabetes. Diabetes claims the lives of approximately 2-5 million people each year. It is predicted that by 2045, this number would have risen to 629 million [46].

Diabetes is influenced by a variety of factors such as height, weight, hereditary factors, insulin, obesity, lack of exercise, living style and bad dieting but the most important aspect to consider is sugar concentration [47,48,49,50,51]. The best way to avoid difficulties is to detect the problem early [52]. When a doctor diagnoses someone with prediabetes, they are advised to make changes to their lifestyle. Diabetes can be avoided by following a healthy diet and exercising regularly [53].

As a result, developing prediction models based on risk factors for the detection of diabetes is critical [54,55,56,57]. Traditional statistical approaches have been suggested as predictors in many studies. However, nowadays, machine learning prediction models are fast replacing the traditional statistical approaches [58,59,60,61]. Machine learning is one of the most essential artificial intelligence elements since it enables the construction of computer systems that can learn from prior experiences without the need for programming in every scenario.

Several researchers have conducted experiments to diagnose diseases using various classification algorithms of machine learning approaches such as J48, SVM, Naive Bayes, Decision Tree, Decision Table, and so on, as studies have shown that machine-learning algorithms [62,63,64,65,66,67] are more effective in diagnosing various diseases. The capacity to manage a vast amount of data, merge data from multiple sources, and integrate background information in the study gives Data Mining [68,69] and Machine learning algorithms their power [70]. Accordingly, in this research, four different machine learning models, namely; Logistic Regression algorithm [71,72,73,74], Support Vector Machine (SVM) algorithm [75,76,77,78], Random Forest algorithm [79,80,81,82] and K Nearest Neighbors (KNN) algorithm [83,84,85,86] are trained for detecting patients with diabetes or patients with the likelihood of having diabetes based on a Pima Indians Diabetes Dataset [87,88,89]. The prediction performance of the four models are compared and the best model is identified and recommended for detecting patients with diabetes or patients with the likelihood of having diabetes.

2.0 Methodology

2.1 The training datasets and feature extraction of the dataset

The Pima Indians Diabetes Dataset with 768 records consisting of 500 healthy patients and 268 diabetic patients (as shown in Figure 1) was used for the machine models training. The dataset was imported in CSV using the Pandas Library. Also, some relevant libraries that are imported includes NumPy, Matplotlib, Pandas, and Sklearn.

The vital features of the dataset are Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI','DiabetesPedigreeFunction', 'Age' (Figure 2). Standardization which is a scaling technique was applied to the dataset to ensure that all the features are maintained on the same scale. The Pandas software was used to generate statistics analysis of the dataset providing information on the data type, missing values, and unique values of datasets, as well as the Quantile statistics, the mean, median, mode, and other descriptive statistics and finally the values, Histograms, and Correlations are frequently used to create strongly correlated variables based on the dataset.

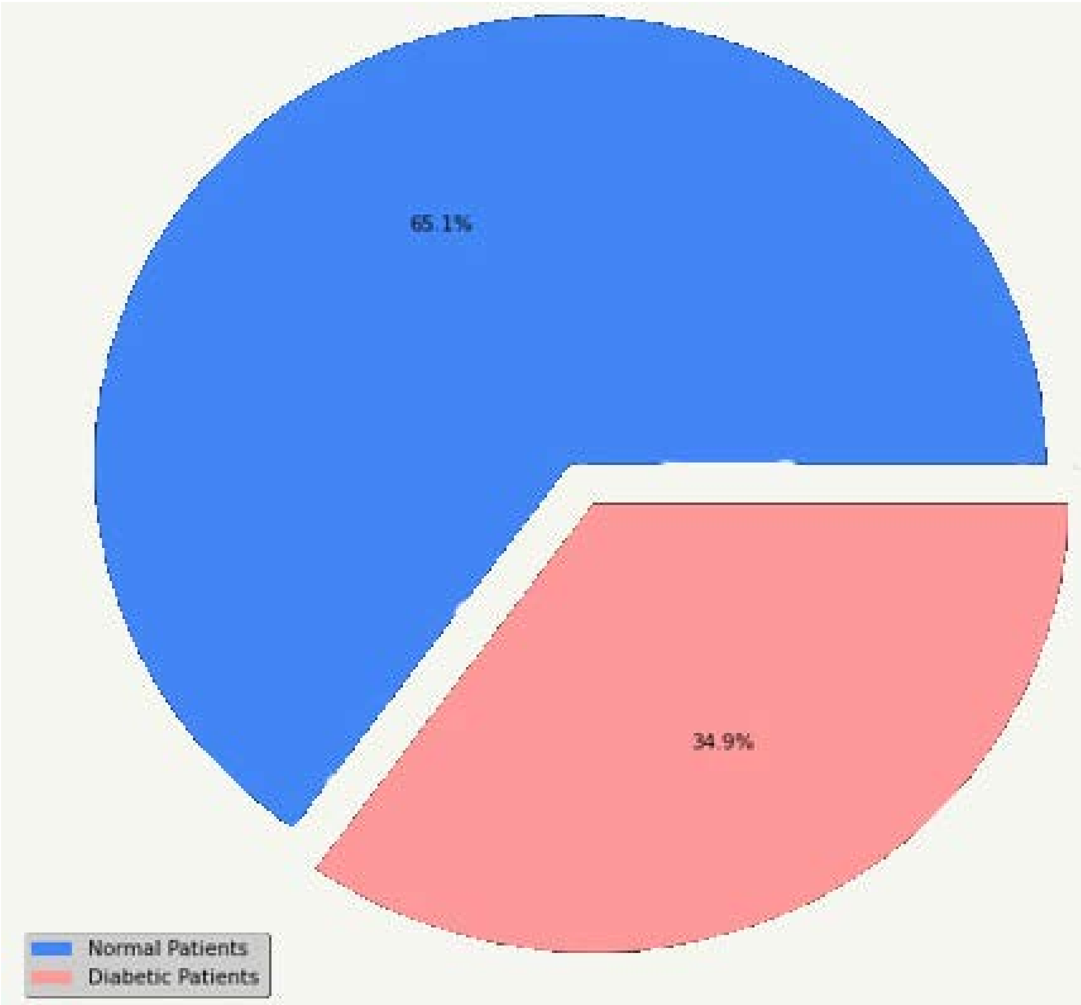

Figure 1: The proportion of the classes in the data sets

```
[ ] # load the dataset
    dataset = pd.read_csv("/content/drive/MyDrive/Mydataset/diabetes.csv")
    dataset.head(10)
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
5	5	116	74	0	0	25.6	0.201	30	0
6	3	78	50	32	88	31.0	0.248	26	1
7	10	115	0	0	0	35.3	0.134	29	0
8	2	197	70	45	543	30.5	0.158	53	1
9	8	125	96	0	0	0.0	0.232	54	1

Figure 2: Features of the dataset

2.2 Model Training

The dataset was split into the training set and the validation set in the ratio of 75:25. The following machine learning algorithms were trained; Logistic Regression algorithm, Support Vector Machine (SVM) algorithm, Random Forest algorithm and K Nearest Neighbors (KNN) algorithm. In order to evaluate the different machine models used, accuracy, recall, precision and f1-score metrics were used.

Logistic Regression Model: The Logistic Regression machine learning algorithm is a linear classifier that predicts the output based on probability. It employs the sigmoid function. The output falls between 0 and 1. If the threshold is 0.5 then values from 0 to 0.49 are false, while values between 0.5 to 1.0 are true. The hyperparameter optimization for the learning algorithm was C: 0.5, penalty: 'l2'. This C: 0.5, penalty: 'l2' was used to control the learning process.

Random Forest Model: The Random Forest Algorithm is based on ensemble learning and it employs multiple decision trees in making its predictions. The Random Forest was trained to detect diabetes using the datasets obtained. The hyperparameter optimization used for the Random Forest model were: criterion: 'entropy' min_samples_split: 30, n_estimators: 110.

Support Vector Machine Model: Support Vector Machine is a poplar supervised learning algorithm. The linear kernel of the support vector machine was used to train the model. The hyperparameters used to train this model was kernel: 'rbf'' .

K-nearest Neighbour (KNN) Model: The K Nearest Neighbour algorithm works on the principle of allocating a weight to each data point, which is referred to as a neighbor. The hyperparameters used to train the KNN were: n_neighbors=9, metric='minkowski', p=2.

3. Results and discussion

The results for several supervised machine learning models, as well as the model's performance utilizing the same dataset are presented.

3.1 The results for the Logistic Regression Model

The result of a fitting procedure of total of five times (5-fold cross validation) was performed on the Logistic regression. In the first procedure**,** the accuracy of logistic regression (train accuracy) on the 90 percent of the data it was trained on was 85.02 accurate while on the accuracy of logistic regression on the 10 percent of the dataset it has not seen (test accuracy) was 65.10 accurate. In the second fold, it had train accuracy of 86.42 and test accuracy of 85.19. The third fold gave a training accuracy of 88.60 and validation accuracy of 86.19. The fourth fold had 81.34 and validation accuracy of 81.34. The fifth training yielded a result of 85.19 accuracy and validation result of 85.19.

The train accuracy gave an average of 80.60 while the test accuracy (validation accuracy) gave an average of 81.00. It gave an average score of 79.57 for precision, average score of 76.89 for Recall and finally an average score of 77.87 was gotten for F1-Score. The confusion matrix of logistic regression shown in Figure 3 shows that the logistic regression model had 112 True Positives , False Positive of 13, False Negatives of 24 and True Negative of 43.

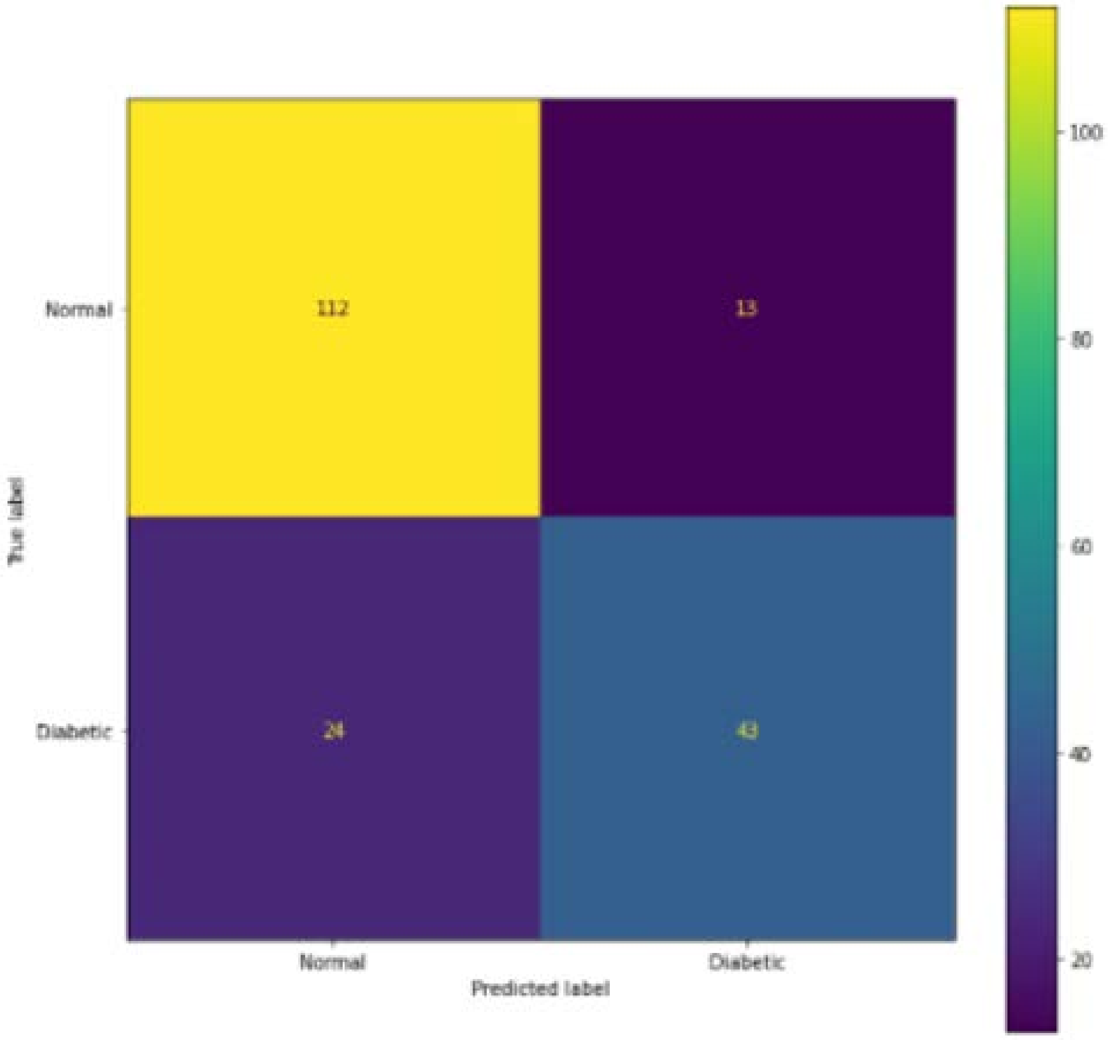

Figure 3: Confusion Matrix for Logistic Regression Model

3.3 The results for the Support Vector Machine model

A 5-fold cross validation was performed on the Support Vector Machine model. The test accuracy (validation accuracy) gave an average of 79.17. The confusion matrix of the Support Vector Machine shown in Figure 4 shows that the Support Vector Machine had 108 True Positives and False Positive of 17. The Support Vector Machine also had 23 False Negatives and True Negative of 44. An average score of 77.29 was generated for precision, average score of 76.04 for Recall and finally an average score of 76.56 was gotten for F1-Score.

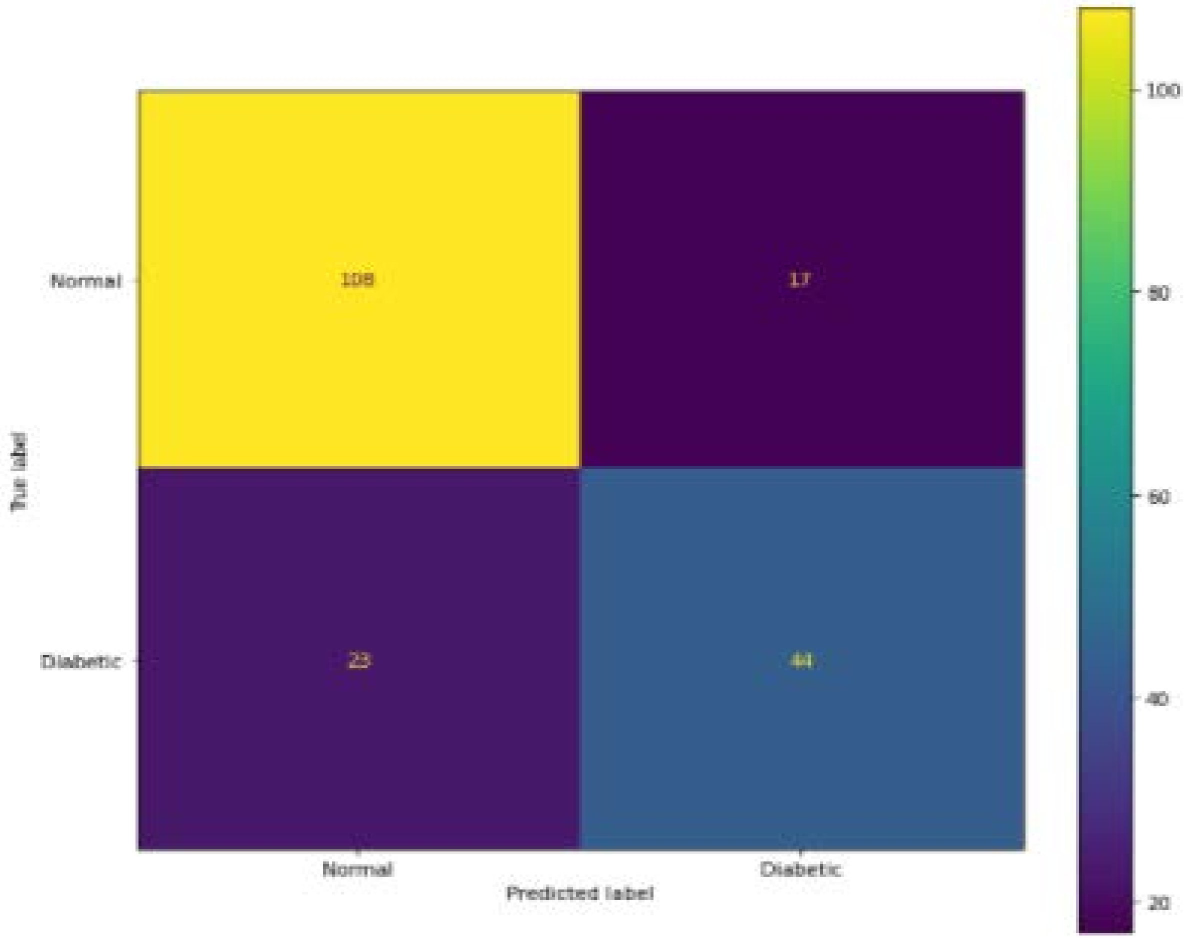

Figure 4: Confusion matrix for Support Vector Machine Model

3.3 The results for the K Nearest Neighbor Model

The K Nearest Neighbor (KNN) model gave an average score of 80.73 for Accuracy, 79.17 for Precision, 77.58 for Recall and 78.23 for F1-score. As shown in the confusion matrix of Figure 5, the K Nearest Neighbour model had a True Positive score of 110, False Positive of 15, False Negative score of 22 and True Negative score of 45.

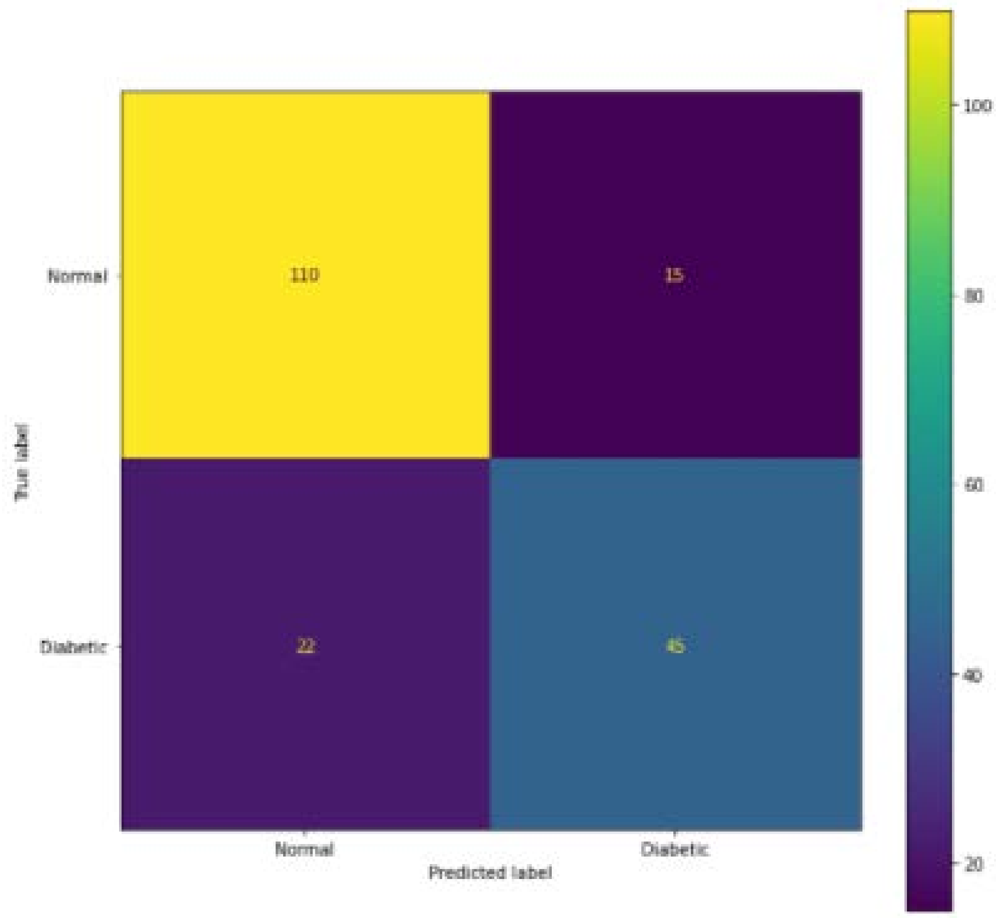

Figure 5: Confusion matrix for k nearest neighbor

3.3 The results for the Random Forest Model

The test accuracy of Random Forest (validation accuracy) gave an average of 81.25. The confusion matrix of Random Forest shown in Figure 6 shows that the Random Forest model had 111 True Positives, 14 False Positives, False negative of 22 and 45 True Negatives. It gave an average score of 79.86 for Precision, average score of 77.98 for Recall and finally an average score of 78.73 was gotten for F1-Score.

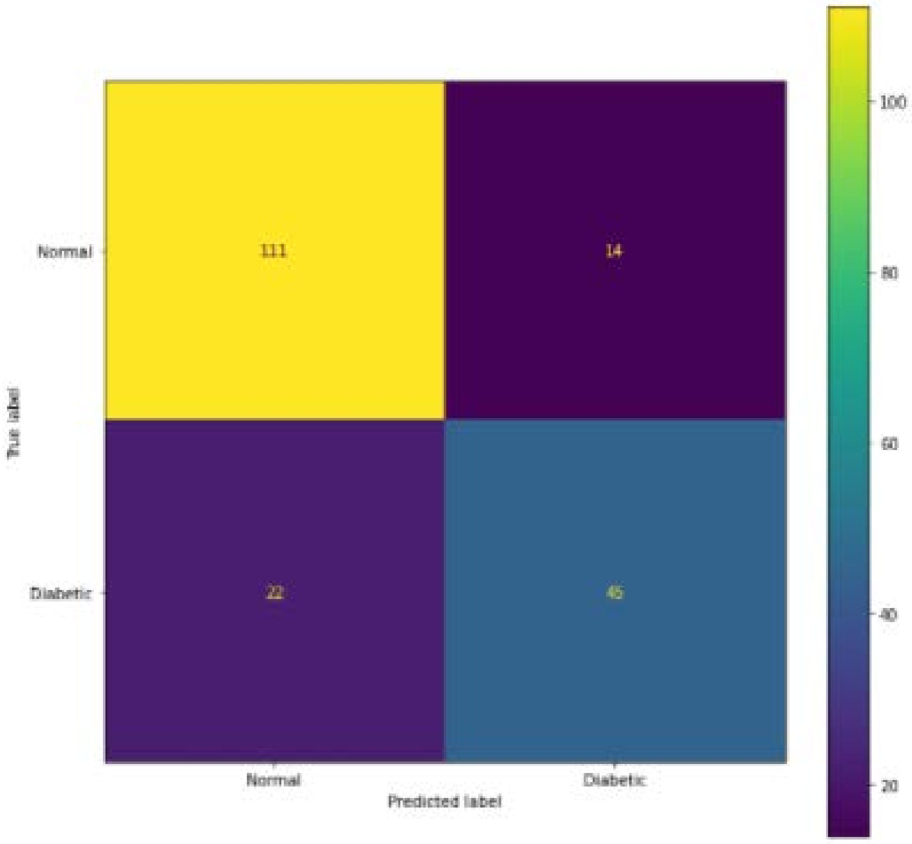

Figure 6 Confusion matrix for random forest

3.4 Comparison of the performance of the four models

The accuracy plot showing Logistic Regression, K Nearest Neighbor, Support Vector Machine and Random Forest is depicted in Figure 7 while . Figure 8 shows the precision plot for the machine learning algorithms used to predict diabetes in this study. The recall chart for the machine learning models used is shown in Figure 9. The weighted average of Precision and Recall is the F1 Score. The graph for the F1-score for the four models are shown in Figure 10. The Summary of validation accuracy, precision, recall, f1-score for each ML model is given in Figure 11.

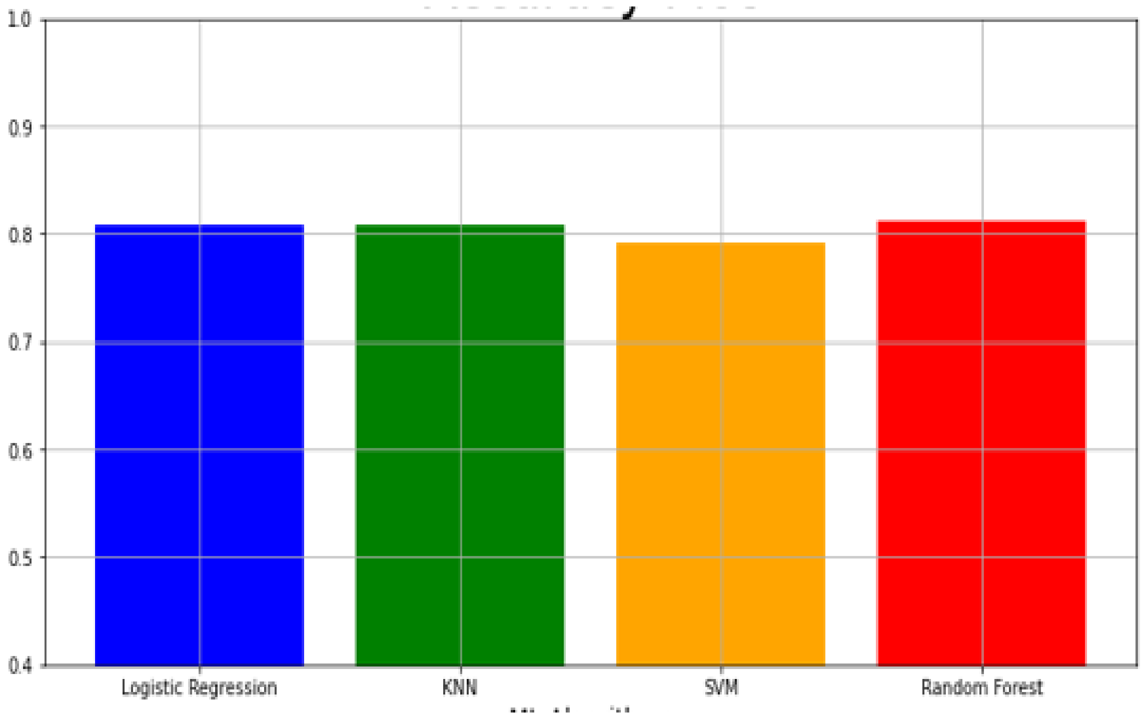

Figure 7 Accuracy chart for each Machine Learning model

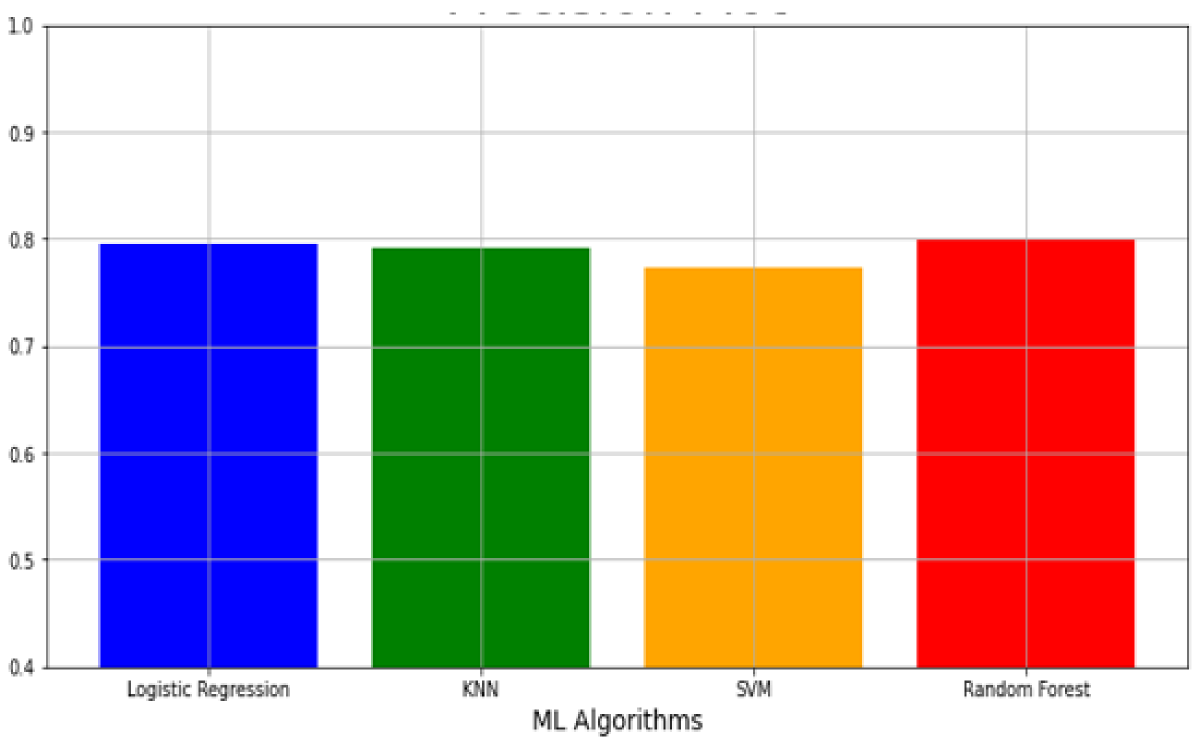

Figure 8 Precision chart for each Machine Learning model employed

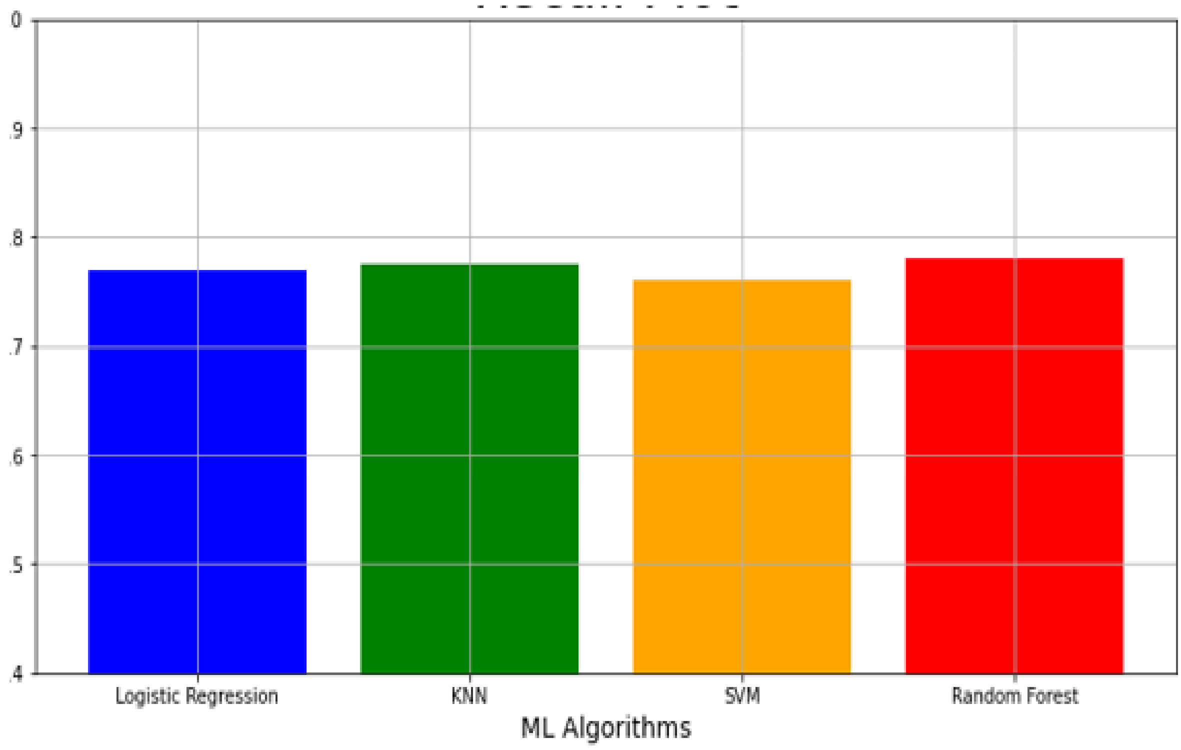

Figure 9 Recall chart for each Machine Learning model employed

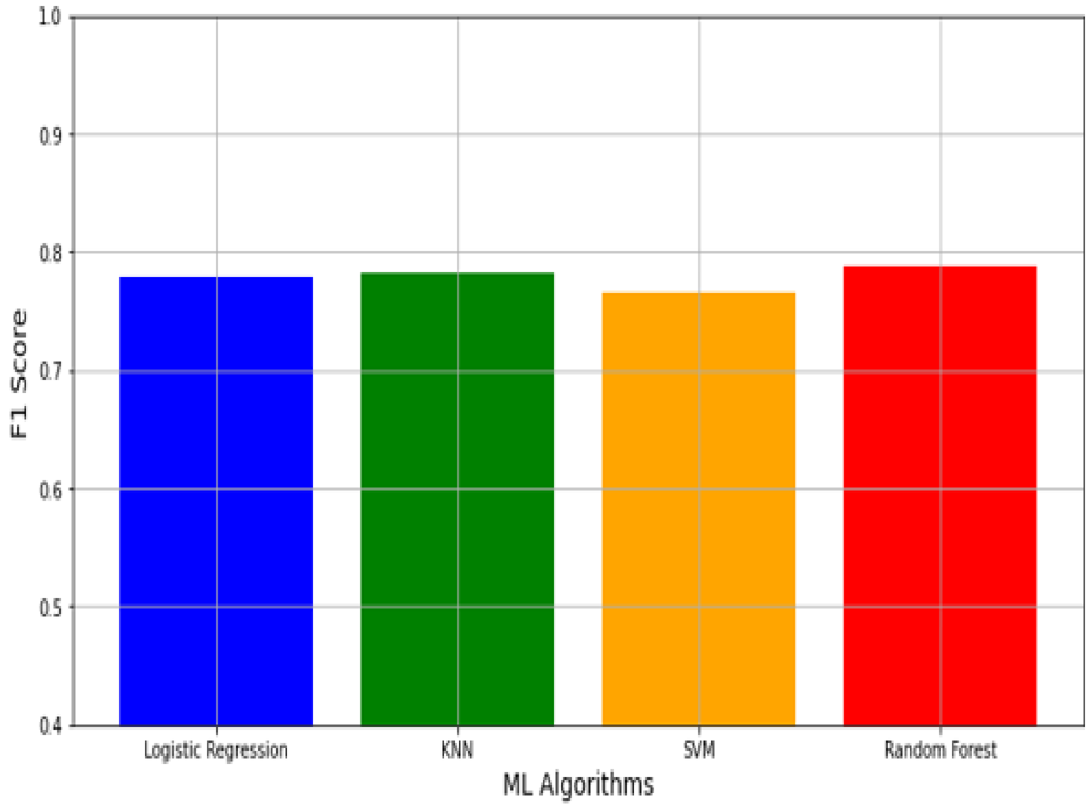

Figure 10 F1-score chart for each machine learning model employed

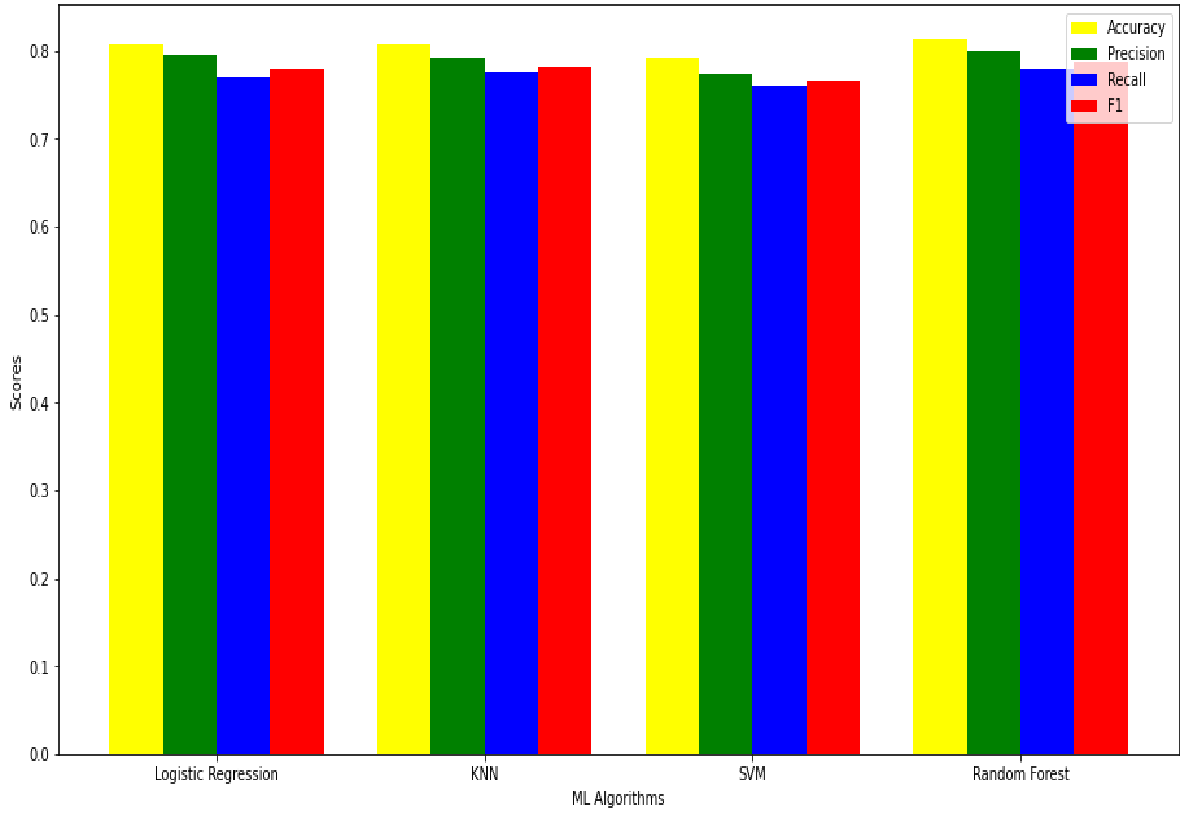

Figure 11 The summary of validation accuracy, precision, recall, f1-score for each ML model

In healthcare problems, the goal is to reduce the number of false negatives. The false negatives are the patients this model predicted to be healthy, but the fact is that these patients have diabetes. This is why the recall metric is highly considered in the healthcare industry. The higher the recall, the lower the false negatives. The machine learning models that gave us the lowest number of false negatives were the K Nearest Neighbour and the Random Forest models.

The accuracy of a machine learning model is a metric for determining which model is the best at recognizing relationships and patterns between variables in a dataset based on the input, or training, data. The train accuracy gave an average of 80.60 while the test accuracy (validation accuracy) gave an average of 81.00. This means that the Logistic Regression performed better on the dataset it was trained on than on the dataset it was validated on. The confusion matrix of Logistic Regression shown in Figure 3 shows that the Logistic Regression was able to predict 112 people with diabetes as people with diabetes (True Positive). It also predicted 13 people with no diabetes as people that have diabetes (False Positives). Logistic Regression was also able to predict 24 diabetic patient as normal patient (False Negatives) and 43 diabetic patient as diabetic patient (True Positive). It gave an average precision score of 79.57 which means that the Logistic Regression learning performance was 79.57 percent, the positive predictions that were missed by Logistic Regression were 76.89 percent and was able to retrieve 77.87 percent of data.

The confusion matrix of Support Vector Machine shown in Figure 4 shows that the Support Vector Machine had a True Positive which means it was able to predict 108 people with no diabetes as people with no diabetes It also predicted 17 people with no diabetes as people with diabetes (False Negative). Support Vector Machine was also able to predict 23 diabetic patients as patient that are not diabetic and 44 non-diabetic patient as non-diabetic patient. An average Precision score was of 77.29 meaning that the percentage at which Support Vector Machine performed was 77.29, average score of 76.04 was given for Recall. This means that correct positive predictions were produced out of all possible positive predictions was 76.04 and finally an average score of 76.56 was gotten for F1-Score which is the capacity for this machine learning to retrieve information from the storage system.

The K Nearest Neighbor (KNN) gave an average score of 80.73 for Accuracy, 79.17 for Precision, 77.58 for Recall and 78.23 for F1-score. K Nearest Neighbor had a True Positive score of 110 which is the cases in KNN predicted yes (they have the disease), actually do have the disease, False Positive of 15 which is KNN model predicted yes, but they don't actually have the disease. (Also known as a "Type I error."), False Negative score of 22 which is the model predicted no, but they actually do have the disease. (Also known as a "Type II error.") and finally, True Negative score of 45 which is KNN predicted no, and they do not have the disease.

The Random Forest model was able to predict 111 people with diabetes as people with diabetes (True Positives). It also predicted 14 people with no diabetes as people that have diabetes (False Positives). It predicted 22 as patients with no diabetes, but they actually do have the disease. (Also known as a "Type II error.") (False Negatives) and 45 non-diabetic patients as non-diabetic patients (True Negatives). It gave an average score of 79.86 for Precision, average score of 77.98 for Recall and finally an average score of 78.73 was gotten for F1-Score. In all, in this research, Random Forest model generated the highest accuracy of 81.25 percent, Precision was 79.57 percent, Recall was 77.98 percent and F1-Score was 78.73 percent.

Conclusion

The use of different machine learning algorithms to predict diabetes is presented. The machine learning algorithms considered includes; Logistic Regression algorithm, Support Vector Machine (SVM) algorithm, Random Forest algorithm and K Nearest Neighbors (KNN) algorithm. The Pima Indians Diabetes Dataset was used for the model training and validation. The dataset had two classes which were normal patients and diabetic patients. The dataset was used to train the different machine learning models to determine which model would be suitable for detecting patients with diabetes or patients with the likelihood of having diabetes. Out of the four different machine learning models considered, the Random Forest algorithm generated the highest accuracy of 81.25 percent, Precision of 79.57 percent, Recall of 77.98 percent and F1-Score of 78.73 percent. Hence, the Random Forest algorithm was recommeneded for for detecting patients with diabetes or patients with the likelihood of having diabetes.

References

1. Anietie Bassey, Simeon Ozumba & Kufre Udofia (2015). An Effective Adaptive Media Play-out Algorithm For Real-time Video Streaming Over Packet Networks. European. *Journal of Basic and Applied Sciences Vol, 2(4).*
2. Simeon, Ozuomba. (2015) "Development Of Seeded Bisection Iteration Method Using Perturbation-Based Mechanism." Development 2.7 (2015). *Journal of Multidisciplinary Engineering Science and Technology (JMEST) Vol. 2 Issue 7, July - 2015*
3. Zion, Idongesit, Simeon Ozuomba, and Philip Asuquo. (2020) "An Overview of Neural Network Architectures for Healthcare." *2020 International Conference in Mathematics, Computer Engineering and Computer Science (ICMCECS)*. IEEE, 2020
4. Sylvester Michael Ekpo, Kingsley M. Udofia, Ozuomba Simeon (2019) Modelling and Simulation of Robust Biometric Fingerprint Recognition Algorithm. *Universal Journal of Applied Science 6(2): 29-38, 2019*
5. Akaninyene Obot, Umoren Mfonobong Anthony, Simeon Ozuomba (2016) A Novel Tabular Form of the Simplex Method for Solving Linear Programming Problems, *International Journal of Computer Science & Network Solutions*, Feb.2016-Volume 4.No.2
6. Ozuomba, Simeon, and Etinamabasiyaka Edet Ekott. (2020). "Design And Implementation Of Microcontroller And Internet Of Things-Based Device Circuit And Programs For Revenue Collection From Commercial Tricycle Operators." *Science and Technology Publishing* (SCI & TECH) Vol. 4 Issue 8, August – 2020
7. Maduka, N. C., Simeon Ozuomba, and E. E. Ekott. . (2020) "Internet of Things-Based Revenue Collection System for Tricycle Vehicle Operators." *2020 International Conference in Mathematics, Computer Engineering and Computer Science (ICMCECS)*. IEEE, 2020.
8. Thompson, E., Simeon, O., & Olusakin, A. (2020). A survey of electronic heartbeat electronics body temperature and blood pressure monitoring system. Journal of Multidisciplinary Engineering Science Studies (JMESS) Vol. 6 Issue 8, August – 2020
9. Simeon, Ozuomba (2015) "Analysis Of Perturbance Coefficient-Based Seeded Secant Iteration Method." *Journal of Multidisciplinary Engineering Science and Technology (JMEST) Vol. 2 Issue 1, January – 2015*
10. Akpan, Nsikak-Abasi Peter, Kufre Udofia, and Simeon Ozuomba (2018). Development and Comparative Study of Least Mean Square-Based Adaptive Filter Algorithms. Development, 3(12). *International Multilingual Journal of Science and Technology (IMJST) Vol. 3 Issue 12, December - 2018*
11. Chikezie, Aneke, Ezenkwu Chinedu Pascal, and Ozuomba Simeon. (2014). "Design and Implementation Of A Microcontroller-Based Keycard." *International Journal of Computational Engineering Research (IJCER) Vol, 04 Issue, 5 May – 2014*
12. Simeon, Ozuomba. (2018) "Sliding Mode Control Synthesis For Autonomous Underwater Vehicles" *Science and Technology Publishing (SCI & TECH*
13. Otumdi, Ogbonna Chima, Kalu Constance, and Ozuomba Simeon (2018). "Design of the Microcontroller Based Fish Dryer." *Journal of Multidisciplinary Engineering Science Studies (JMESS) Vol. 4 Issue 11, November - 201*
14. Njoku, Felix A., Ozuomba Simeon, and Fina Otosi Faithpraise (2019). Development Of Fuzzy Inference System (FIS) For Detection Of Outliers In Data Streams Of Wireless Sensor Networks. *International Multilingual Journal of Science and Technology (IMJST) Vol. 4 Issue 10, October - 2019*
15. Ozuomba, Simeon, Ekaette Ifiok Archibong, and Etinamabasiyaka Edet Ekott (2020). Development Of Microcontroller-Based Tricycle Tracking Using Gps And Gsm Modules. *Journal of Multidisciplinary Engineering Science and Technology (JMEST) Vol. 7 Issue 1, January - 2020*
16. Kalu, C., Ozuomba, Simeon. & Udofia, K. (2015). Web-based map mashup application for participatory wireless network signal strength mapping and customer support services. *European Journal of Engineering and Technology, 3 (8)*, 30-43.

17. Gordon, O., Ozuomba, Simeon. & Ogbajie, I. (2015). Development of educate: a social network web application for e-learning in the tertiary institution. *European Journal of Basic and Applied Sciences*, *2 (4)*, 33-54.
18. Ozuomba, Simeon, Kalu, C., & Anthony, U. M. (2015). Map Mashup Application And Facilitated Volunteered Web-Based Information System For Business Directory In Akwa Ibom State. *European Journal of Engineering and Technology Vol*, *3*(9).
19. Akpasam Joseph Ekanem, Simeon Ozuomba, Afolayan J. Jimoh (2017) Development of Students Result Management System: A case study of University of Uyo. *Mathematical and Software Engineering*, Vol. 3, No. 1 (2017), 26-42.
20. Ozuomba Simeon , S.T Wara, C. Kalu and S.O Oboma (2006) ; *Computer Aided design of the magnetic circuit of a three phase power transformer, Ife Journal of Technology Vol.15, No. 2 , November 2006 , PP 99 – 108*
21. Kalu, C., Ezenugu, I. A. & Ozuomba, Simeon. (2015). Development of matlab-based software for peak load estimation and forecasting: a case study of faculty of engineering, Imo State University Owerri, Imo state, Nigeria. *European Journal of Engineering and Technology*, *3 (8)*, 20-29.
22. Stephen, B. U., Ozuomba, Simeon, & Eyibo, I. E. (2018). Development of Reward Mechanism for Proxy Marketers Engaged in E-Commerce Platforms. *European Journal of Engineering and Technology Research, 3(10), 45-52.*
23. Eyibo, I. E., Ozuomba, Simeon, & Stephen, B. U. A. (2018). DEVELOPMENT OF TRUST MODEL FOR PROXY MARKETERS ENGAGED IN E-COMMERCE PLATFORMS. *European Journal of Engineering and Technology Vol*, *6(4).*
24. Ozuomba, Simeon, Constant Kalu, and Akpasam Joseph. (2018). Development of Facilitated Participatory Spatial Information System for Selected Urban Management Services. *Review of Computer Engineering Research, 5(2), 31-48.*
25. Kalu, Constance, Simeon Ozuomba, and Sylvester Isreal Umana. (2018). Development of Mechanism for Handling Conflicts and Constraints in University Timetable Management System. *Communications on Applied Electronics (CAE) 7(24).*
26. Ekanem, Mark Sunday, and Simeon Ozuomba. (2018). ONTOLOGY DEVELOPMENT FOR PEDAGOGIC CONTENT INFORMATICS. *European Journal of Engineering and Technology Vol, 6(4).*
27. Bassey, M. U., Ozuomba, Simeon, & Stephen, B. U. A. (2019). DEVELOPMENT OF A FACILITATED CROWD-DRIVEN ONLINE PROFIT-MAKING SYSTEM. *European Journal of Engineering and Technology Vol, 7(5).*
28. Ibanga, Jude, and Ozuomba Simeon, Obot, Akaniyene. B. (2020) "Development of Web-Based Learning Object Management System." Development 7, no. 3 (2020). *Journal of Multidisciplinary Engineering Science and Technology (JMEST) Vol. 7 Issue 3, March - 2020*
29. Simeon Ozuomba , Gloria A. Chukwudebe , Felix K. Opara and Michael Ndinechi (2014)Chapter 8: Social Networking Technology: A Frontier Of Communication For Development In The Developing Countries Of Africa . *In Green Technology Applications for Enterprise and Academic Innovation (Chapter 8).* IGI Global, Hershey, PA 17033-1240, USA
30. Ezenkwu, Chinedu Pascal, Simeon Ozuomba, and Constance Kalu. (2013). "Strategies for improving community policing in Nigeria through Community Informatics Social Network." *2013 IEEE International Conference on Emerging & Sustainable Technologies for Power & ICT in a Developing Society (NIGERCON)*. IEEE, 2013.
31. Nicholas A. E., Simeon O., Constance K. (2013) Community informatics social e-learning network: a case study of Nigeria *Software Engineering 2013; 1(3): 13-21*
32. Effiong, Clement, Simeon Ozuomba, and Udeme John Edet (2016). Long-Term Peak Load Estimate and Forecast: A Case Study of Uyo Transmission Substation, Akwa Ibom State, Nigeria. *Science Journal of Energy Engineering 4(6), 85-89.*
33. Stephen, Bliss Utibe-Abasi, Ozuomba Simeon, and Sam Bassey Asuquo. (2018) "Statistical Modeling Of The Yearly Residential Energy Demand In Nigeria." *Journal of Multidisciplinary Engineering Science Studies* (JMESS) Vol. 4 Issue 6, June – 2018
34. Uko, Sampson Sampson, Ozuomba Simeon, and Ikpe Joseph Daniel (2019). Adaptive neuro-fuzzy inference system (ANFIS) model for forecasting and predicting industrial electricity consumption in Nigeria. *Advances in Energy and Power*, 6(3), 23-36.
35. Effiong, Clement, Ozuomba Simeon, and Fina Otosi Faithpraise (2020). "Modelling And Forecasting Peak Load Demand In Uyo Metropolis Using Artificial Neural Network Technique." *Journal of Multidisciplinary Engineering Science and Technology (JMEST) Vol. 7 Issue 3, March – 2020*
36. Eti-Ini Robson Akpan, Ozuomba Simeon, Sam Bassey Asuquo (2020). POWER FLOW ANALYSIS USING INTERLINE POWER FLOW CONTROLLER *Journal of Multidisciplinary Engineering Science and Technology (JMEST) Vol. 7 Issue 5, May – 2020*
37. Ozuomba, Simeon, Victor Akpaiya Udom & Jude Ibanga. (2018). Iterative Newton-Raphson-Based Impedance Method For Fault Distance Detection On Transmission Line. Education, 2020. *International Multilingual Journal of Science and Technology (IMJST) Vol. 5 Issue 5, May - 2020*
38. Chinedu Pascal Ezenkwu , Simeon Ozuomba , Constance Kalu (2015) , *Application of k-Means Algorithm for efficient Customer Segmentation: A strategy for targeted customer services. (IJARAI)*

International Journal of Advanced Research in Artificial Intelligence, Vol. 4, No.10, 2015

39. Inyang, Imeobong Frank, Simeon Ozuomba, and Chinedu Pascal Ezenkwu.(2017) "Comparative analysis of Mechanisms for Categorization and Moderation of User Generated Text Contents on a Social E-Governance Forum." *Mathematical and Software Engineering* 3.1 (2017): 78-86.
40. Mathew-Emmanuel, Eze Chinenye, Simeon Ozuomba, and Constance Kalu. (2017) "Preliminary Context Analysis of Social Network Web Application for Combating HIV/AIDS Stigmatization." *Mathematical and Software Engineering* 3.1 (2017): 99-107
41. Ezeonwumelu, P., Ozuomba, Simeon. & Kalu, C. (2015). Development of swim lane workflow process map for enterprise workflow management information system (WFMIS): a case study of comsystem computer and telecommunication ltd (CCTL) EKET. *European Journal of Engineering and Technology, 3 (9)*, 1-13.
42. Kandhasamy, J. P., & Balamurali, S. J. P. C. S. (2015). Performance analysis of classifier models to predict diabetes mellitus. *Procedia Computer Science*, *47*, 45-51.
43. Forouhi, N. G., & Wareham, N. J. (2019). Epidemiology of diabetes. *Medicine*, *47*(1), 22-27.
44. Zhu, Y., & Zhang, C. (2016). Prevalence of gestational diabetes and risk of progression to type 2 diabetes: a global perspective. *Current diabetes reports*, *16*(1), 1-11.
45. Chaudhary, N., & Tyagi, N. (2018). Diabetes mellitus: An Overview. *International Journal of Research and Development in Pharmacy & Life Sciences*, *7*(4), 3030-3033.
46. World Health Organization. (2013). *Diagnostic criteria and classification of hyperglycaemia first detected in pregnancy* (No. WHO/NMH/MND/13.2). World Health Organization.
47. Kalyankar, G. D., Poojara, S. R., & Dharwadkar, N. V. (2017, February). Predictive analysis of diabetic patient data using machine learning and Hadoop. In *2017 international conference on I-SMAC (IoT in social, mobile, analytics and cloud)(I-SMAC)* (pp. 619-624). IEEE.
48. Hauner, H. (2017). Obesity and diabetes. *Textbook of diabetes*, 215-228.
49. Mozaffarian, D. (2016). Dietary and policy priorities for cardiovascular disease, diabetes, and obesity: a comprehensive review. *Circulation*, *133*(2), 187-225.
50. Ornoy, A., Reece, E. A., Pavlinkova, G., Kappen, C., & Miller, R. K. (2015). Effect of maternal diabetes on the embryo, fetus, and children: congenital anomalies, genetic and epigenetic changes and developmental outcomes. *Birth Defects Research Part C: Embryo Today: Reviews*, *105*(1), 53-72.
51. Fruh, S. M. (2017). Obesity: Risk factors, complications, and strategies for sustainable long-term weight management. *Journal of the American Association of Nurse Practitioners*, *29*(S1), S3-S14.
52. Rahim, S. A., & AL-Murshidi, M. M. H. (2020). Obesity: Causes and Consequences. *Journal of University of Babylon for Pure and Applied Sciences*, *28*(3), 365-371.
53. Vijayan, V. V., & Anjali, C. (2015, December). Prediction and diagnosis of diabetes mellitus—A machine learning approach. In *2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS)* (pp. 122-127). IEEE.
54. Kaveeshwar, S. A., & Cornwall, J. (2014). The current state of diabetes mellitus in India. *The Australasian medical journal*, *7*(1), 45.
55. Kaur, H., & Kumari, V. (2020). Predictive modelling and analytics for diabetes using a machine learning approach. *Applied computing and informatics*.
56. Wu, H., Yang, S., Huang, Z., He, J., & Wang, X. (2018). Type 2 diabetes mellitus prediction model based on data mining. *Informatics in Medicine Unlocked*, *10*, 100-107.
57. Razavian, N., Blecker, S., Schmidt, A. M., Smith-McLallen, A., Nigam, S., & Sontag, D. (2015). Population-level prediction of type 2 diabetes from claims data and analysis of risk factors. *Big Data*, *3*(4), 277-287.
58. McGurnaghan, S. J., Weir, A., Bishop, J., Kennedy, S., Blackbourn, L. A., McAllister, D. A., ... & Health Protection Study Group. (2021). Risks of and risk factors for COVID-19 disease in people with diabetes: a cohort study of the total population of Scotland. *The lancet diabetes & endocrinology*, *9*(2), 82-93.
59. Shameer, K., Johnson, K. W., Glicksberg, B. S., Dudley, J. T., & Sengupta, P. P. (2018). Machine learning in cardiovascular medicine: are we there yet?. *Heart*, *104*(14), 1156-1164.
60. Luo, W., Phung, D., Tran, T., Gupta, S., Rana, S., Karmakar, C., ... & Berk, M. (2016). Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. *Journal of medical Internet research*, *18*(12), e5870.
61. Schmidt, J., Marques, M. R., Botti, S., & Marques, M. A. (2019). Recent advances and applications of machine learning in solid-state materials science. *npj Computational Materials*, *5*(1), 1-36.
62. Lantz, B. (2019). *Machine learning with R: expert techniques for predictive modeling*. Packt publishing ltd.
63. Kavakiotis, I., & Tsave, O. (2017). A. S, N. Maglaveras, I. Vlahavas, I. Chouvarda. *Machine learning and data mining methods in diabetes research computational and structural. Biotechnol J*, *15*, 104-116.
64. Mashudi, N. A., Ahmad, N., & Noor, N. M. (2021). Classification of adult autistic spectrum disorder using machine learning approach. *IAES International Journal of Artificial Intelligence*, *10*(3), 743.

65. Marimuthu, M., Abinaya, M., Hariesh, K. S., Madhankumar, K., & Pavithra, V. (2018). A review on heart disease prediction using machine learning and data analytics approach. *International Journal of Computer Applications*, *181*(18), 20-25.
66. Hosseinzadeh, M., Koohpayehzadeh, J., Bali, A. O., Asghari, P., Souri, A., Mazaherinezhad, A., ... & Rawassizadeh, R. (2021). A diagnostic prediction model for chronic kidney disease in internet of things platform. *Multimedia Tools and Applications*, *80*(11), 16933-16950.
67. Learning, M. (2017). Heart disease diagnosis and prediction using machine learning and data mining techniques: a review. *Advances in Computational Sciences and Technology*, *10*(7), 2137-2159.
68. Diwakar, M., Tripathi, A., Joshi, K., Memoria, M., & Singh, P. (2021). Latest trends on heart disease prediction using machine learning and image fusion. *Materials Today: Proceedings*, *37*, 3213-3218.
69. Kanchan, B. D., & Kishor, M. M. (2016, December). Study of machine learning algorithms for special disease prediction using principal of component analysis. In *2016 international conference on global trends in signal processing, information computing and communication (ICGTSPICC)* (pp. 5-10). IEEE.
70. Kumar, P. S., & Umatejaswi, V. (2017). Diagnosing diabetes using data mining techniques. *International Journal of Scientific and Research Publications*, *7*(6), 705-709.
71. Nguyen, G., Dlugolinsky, S., Bobák, M., Tran, V., López García, Á., Heredia, I., ... & Hluchý, L. (2019). Machine learning and deep learning frameworks and libraries for large-scale data mining: a survey. *Artificial Intelligence Review*, *52*(1), 77-124.
72. Fatima, M., & Pasha, M. (2017). Survey of machine learning algorithms for disease diagnostic. *Journal of Intelligent Learning Systems and Applications*, *9*(01), 1.
73. Xiao, T., Zhang, P., Zhang, Y., Li, D., & Shen, J. (2021). A research on the application of college students' physique data mining based on logistic regression algorithm. *ASP Transactions on Computers*, *1*(2), 12-18.
74. De Caigny, A., Coussement, K., & De Bock, K. W. (2018). A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. *European Journal of Operational Research*, *269*(2), 760-772.
75. Shah, K., Patel, H., Sanghvi, D., & Shah, M. (2020). A comparative analysis of logistic regression, random forest and KNN models for the text classification. *Augmented Human Research*, *5*(1), 1-16.
76. Huang, S., Cai, N., Pacheco, P. P., Narrandes, S., Wang, Y., & Xu, W. (2018). Applications of support vector machine (SVM) learning in cancer genomics. *Cancer genomics & proteomics*, *15*(1), 41-51.
77. Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., & Lopez, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, *408*, 189-215.
78. Pisner, D. A., & Schnyer, D. M. (2020). Support vector machine. In *Machine learning* (pp. 101-121). Academic Press.
79. Suthaharan, S. (2016). Support vector machine. In *Machine learning models and algorithms for big data classification* (pp. 207-235). Springer, Boston, MA.
80. Biau, G., & Scornet, E. (2016). A random forest guided tour. *Test*, *25*(2), 197-227.
81. Sarica, A., Cerasa, A., & Quattrone, A. (2017). Random forest algorithm for the classification of neuroimaging data in Alzheimer's disease: a systematic review. *Frontiers in aging neuroscience*, *9*, 329.
82. Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. *The Stata Journal*, *20*(1), 3-29.
83. Lin, W., Wu, Z., Lin, L., Wen, A., & Li, J. (2017). An ensemble random forest algorithm for insurance big data analysis. *Ieee access*, *5*, 16568-16575.
84. Abu Alfeilat, H. A., Hassanat, A. B., Lasassmeh, O., Tarawneh, A. S., Alhasanat, M. B., Eyal Salman, H. S., & Prasath, V. S. (2019). Effects of distance measure choice on k-nearest neighbor classifier performance: a review. *Big data*, *7*(4), 221-248.
85. Yu, Z., Chen, H., Liu, J., You, J., Leung, H., & Han, G. (2015). Hybrid $ k $-nearest neighbor classifier. *IEEE transactions on cybernetics*, *46*(6), 1263-1275.
86. Chomboon, K., Chujai, P., Teerarassamee, P., Kerdprasop, K., & Kerdprasop, N. (2015, March). An empirical study of distance metrics for k-nearest neighbor algorithm. In *Proceedings of the 3rd international conference on industrial application engineering* (pp. 280-285).
87. Cunningham, P., & Delany, S. J. (2021). K-nearest neighbour classifiers-a tutorial. *ACM Computing Surveys (CSUR)*, *54*(6), 1-25.
88. Vaishali, R., Sasikala, R., Ramasubbareddy, S., Remya, S., & Nalluri, S. (2017, October). Genetic algorithm based feature selection and MOE Fuzzy classification algorithm on Pima Indians Diabetes dataset. In *2017 international conference on computing networking and informatics (ICCNI)* (pp. 1-5). IEEE.
89. Sankar Ganesh, P. V., & Sripriya, P. (2020). A comparative review of prediction methods for pima indians diabetes dataset. In *International Conference On Computational Vision and Bio Inspired Computing* (pp. 735-750). Springer, Cham.