The current issue and full text archive of this journal is available on Emerald Insight at:
https://www.emerald.com/insight/2634-1964.htm

Classification models for likelihood prediction of diabetes at early stage using feature selection

Oladosu Oyebisi Oladimeji
Department of Computer Science, University of Ibadan, Ibadan, Nigeria and Department of Computer Science and Information Technology, Bowen University, Iwo, Nigeria
Abimbola Oladimeji
Department of Chemistry, University of Ibadan, Ibadan, Nigeria, and
Olayanju Oladimeji
Department of Computer Science and Information Technology, Bowen University, Iwo, Nigeria

Received 29 January 2021
Revised 9 February 2021
6 March 2021
16 April 2021
Accepted 7 May 2021

Abstract

Purpose – Diabetes is one of the life-threatening chronic diseases, which is already affecting 422m people globally based on (World Health Organization) WHO report as at 2018. This costs individuals, government and groups a whole lot; right from its diagnosis stage to the treatment stage. The reason for this cost, among others, is that it is a long-term treatment disease. This disease is likely to continue to affect more people because of its long asymptotic phase, which makes its early detection not feasible.
Design/methodology/approach – In this study, the authors have presented machine learning models with feature selection, which can detect diabetes disease at its early stage. Also, the models presented are not costly and available to everyone, including those in the remote areas.
Findings – The study result shows that feature selection helps in getting better model, as it prevents overfitting and removes redundant data. Hence, the study result when compared with previous research shows the better result has been achieved, after it was evaluated based on metrics such as F-measure, Precision-Recall curve and Receiver Operating Characteristic Area Under Curve. This discovery has the potential to impact on clinical practice, when health workers aim at diagnosing diabetes disease at its early stage.
Originality/value – This study has not been published anywhere else.

Keywords Diabetes, Machine learning, Feature selection, Data mining, Classification algorithms
Paper type Research paper

Introduction

Chronic diseases are known to affect the way of life negatively [1] and have financial impact on individuals, governments and groups while tackling the chronic diseases [2, 3]. Diabetes mellitus is one such example categorized by high levels of blood glucose [4] leading to severe damage of the heart, blood vessels, eyes, kidneys and nerves [5, 6]. Diabetes is a growing health challenge of this era irrespective of geographic, racial and ethnic background [7, 8]. As recorded by World Health Organization (WHO), the number of diabetic patients globally has rapidly grown from 1m in 1980 to 422m in 2014, and it is increasing steadily especially in low and middle-income countries [9]. According to International Diabetes Federation (IDF), about

Applied Computing and Informatics
Emerald Publishing Limited
e-ISSN: 2210-8327
p-ISSN: 2634-1964
DOI 10.1108/ACI-01-2021-0022

463m people between the ages of 20 and 79 years are diabetes patients, and it is estimated that by 2045, the figure would have increased to 700m [10, 11]. It is also referred to as one of the major causes of death with annual death toll of 1.6m [5].

Researchers have divided diabetes mellitus into three major types: Type 1 diabetes is a serious and ceaseless illness [12] wherein the immune system wrongly attacks the pancreatic beta cells, thus causing insufficient or no insulin production. Type 2 diabetes mellitus is caused when the body uses insulin ineffectively while gestational diabetes occurs only during pregnancy [13] as a result of hormonal changes [14].

While heavy financial burden of diagnosing and managing the disease is experienced by government, individuals and groups and the prevalence rate is growing, a study [11] established that in 2019, diabetes caused at least 760bn dollars, thus demanding to find ways and means to eradicate or reduce this burden to the barest minimum. In this regard, one of the issues is identifying the risk of diabetes at its early phase [15] as early diagnosis and use of suitable therapeutic management support patient compliance and reduce the overweight expenses.

Established methods of diagnosing diabetes are Oral Glucose Tolerance Test (OGTT) and HbA1c test once the patients develop certain type of symptoms and need resources and time. In addition, such resources are not available at distant places [8]. Once diagnosed, the treatment process is long-term and expensive. Therefore, the earlier it is detected, the better it is managed in terms of disease and the expenses [8, 16].

In recent years, machine learning has been employed in prediction of most, if not all, of human activities and natural phenomenon and the health sector is not exempted. Machine learning, a branch of artificial intelligence uses scientific algorithms and models that computer system uses to perform tasks efficiently, without using explicit instructions, but depending on patterns and inference instead [17–19]. A lot of data have been gathered by the healthcare industry [20, 21], which will be of great help in bringing insight into big data for prediction, diagnostic, disease prevention and policymaking purposes through machine learning and data analytics [13, 22].

Authors have proposed various models for the diagnosis of diabetes [1], employing dataset of 768 female subjects with nine attributes such as glucose level, blood pressure level, number of times pregnant, skin thickness, insulin, diabetic pedigree function, age, body mass index (BMI), and outcome to create model for the prediction of diabetes using artificial neural network (ANN), which gave 75.7% accuracy; random forests, which gave 74.7% accuracy; and K-means clustering, which gave 73.4% accuracy; all coupled with feature selection, which will aid health workers with treatment decisions.

A similar study compared five machine learning algorithms to predict diabetes [18] using dataset with 11 parameters after feature selection on Support Vector Machine, Random Forest, Naïve Bayes, Decision Tree and K-nearest Neighbor. The study revealed that Naïve Bayes performed the best. Using another approach, easily accessible and cost-effective model was used for early detection of symptom of this deadly ailment [8] employing dataset of 520 subjects with 16 attributes (symptoms), on different classification algorithms and reported that random forest performed optimally.

One of the major challenges of machine learning is high dimensionality of the dataset [18, 23] requiring a large memory due to analysis of many features, which leads to overfitting. Therefore, the weighting features reduce redundant data and processing time, thereby improving the performance of the algorithm [24–26].

The present study intends to address research questions: Is it possible to have an optimal model that will predict the likelihood of diabetes based on its symptoms? Can less costly system be developed to diagnose diabetes at early stage?

Similarly, along this line, the main objective of this study is to predict the likelihood of diabetes at early stage using feature selection, which eliminates the unnecessary and unimportant features in the dataset [20, 27–30] in order to obtain better results compared to previous research studies such as [15].

Materials and methods

The proposed methodology is hereby described as follows:

(1) Preprocessing (data manipulation).

(2) Feature selection: This is done by using four different algorithms coupled with ranker search method.

(3) Classification – classifiers were tested: KNN, J48, Naïve Bayes, random forests.

(4) Evaluation of results – based on confusion matrix (accuracy, ROC AUC, PR AUC and *F*-measure metrics).

The methodology for this study was formulated using Waikato Environment for Knowledge Analysis (WEKA) software that is an open-source software for machine learning that was developed at the University of Waikato. The dataset that was used to pinpoint this research was gotten from University of California, Irvine (UCI) Machine Learning Repository [31], which is a clinical record of symptoms that may cause diabetes; dataset by [8] was loaded into WEKA. In order to obtain better result, feature selection was applied for selecting the attributes to be used for the classification task. In this study, cross-validation and percentage split methods were used. Random forests, J48, Naïve Bayes and K-Nearest Neighbor (KNN)(IBK) algorithms were used for this study.

Data description and statistical analysis

The dataset contains records of diabetes-related symptoms of 520 individuals. It entails records of people including the symptoms that may cause diabetes, which was collected from Sylhet Diabetes Hospital of Sylhet in Bangladesh. The dataset was created from a direct questionnaire to people who recently have become diabetic or who are still nondiabetic but having some symptoms that may cause diabetes. The dataset contains 17 attributes, which contain information about diabetes symptoms. The full description of the dataset is available at (https://github.com/OladosuO).

In the present analysis, the data of subjects aged ≤90 and exhibiting symptoms was included while the ones refusing the prior informed consent were not included. The WEKA software is used for this study (version 3.8.3) based on Java runtime version −1.8.0_221-b11.

Hypothesis testing

The aim of hypothesis testing is to evaluate whether the attributes of the population suggest prevalence of diabetes or otherwise assuming that the result is valid when the null hypothesis is accepted at p-value > alpha (0.05). Table 1 shows the result of the hypothesis using t-test method.

H0. The attribute value affects the outcome of the diagnosis.

Data preprocessing

First step in data mining is data cleaning, which involves data preprocessing processes [32, 33]. The data preprocessing had already been done through the handling of the missing values using the technique of ignoring the tuples with incomplete values by previous research on this subject matter [8]. In this process, it was discovered that the dataset was skewed (imbalanced). The positive class has 320 instances, while the negative class has 200 instances; the Synthetic Minority Oversampling Technique (SMOTE), which is an oversampling method [34], was used to alleviate the class imbalance problem.

Table 1. Summary of evaluators' ranking and hypothesis testing of each attribute of the dataset

Attribute	SU	IG	GR	CO	*p*-values (hypothesis)
Polydipsia	0.424879	0.40533	0.442065	0.705	0.004455369
Polyuria	0.413827	0.402553	0.421771	0.7101	1.50198E-63
Gender	0.215148	0.199123	0.231551	0.5107	2.01595E-33
Sudden weight loss	0.207971	0.197194	0.217825	0.511	0.804535287
Partial paresis	0.193382	0.185016	0.20058	0.4965	4.05383E-27
Irritability	0.110954	0.093769	0.13412	0.3539	1.60303E-16
Polyphagia	0.090316	0.088601	0.091252	0.348	2.87562E-16
Age	0.074924	0.119532	0.054341	0.1109	0.062645798
Alopecia	0.057503	0.056315	0.0582	0.2758	2.27378E-09
Visual blurring	0.04998	0.04909	0.050435	0.2601	3.75617E-09
Weakness	0.045398	0.045074	0.045317	0.2481	0.627622771
Genital thrush	0.017275	0.014724	0.020635	0.1432	0.008781571
Muscle stiffness	0.006255	0.006078	0.006382	0.0919	0.004473367
Obesity	0.004088	0.003334	0.005208	0.0682	0.089793074
Delayed healing	0.001679	0.001665	0.001677	0.0481	0.28406391
Itching	0.000252	0.001665	0.000251	0.0186	0.760806888

Feature selection

The feature selection process involves understanding the datasets and selecting the attributes, which will produce the essential data required to infer the knowledge been sought for. This is also referred to as feature selection, which is a procedure of recognizing the subset of data from large dimension of data [35].

Attributes that contribute more to the development of the model were derived with the use of SymmetricalUncertAttributeEvaluator (SU), InfoGainAttributeEvaluator (IG), GainRatioAttributeEvaluator (GR), CorrelationAttributeEvaluator (CO) coupled with ranker search method. Table 1 presents a summary of the attributes and how the algorithms ranked them.

Obesity, delayed healing and itching are found to be redundant attributes and not contributing to the model based on how they were ranked by the evaluators; hence, their removal from the classification task. The extracted features are polydipsia, polyuria, gender, sudden weight loss, partial paresis, irritability, polyphagia, age, alopecia, visual blurring, weakness, genital thrush and muscle stiffness.

Classification

After the data preprocessing and selection processes were completed, random forest, Naïve Bayes, J48 and K-nearest neighbor algorithms were applied using WEKA. It is a tested and trusted open-source software for machine learning developed at the University of Waikato, New Zealand [36]. Cross-validation and percentage split were used as the test mode option with 10 as the number of folds and 80% test split for cross-validation and percentage split respectively. Class attribute was set as the target to be predicted for the classification in which the model will present the output variable, which is expected to state the diagnosis outcome whether positive or negative. This process was done five times coupled with changing the random seed starting from 1 to 5 for the process for validation purposes.

Results and discussion

The results are presented for the likelihood of diabetes based on the dataset before applying the algorithms using cross-validation method followed by the results that were obtained using the percentage split method. Then we evaluated the results reported in [8, 11]. The algorithms were implemented as discussed in the previous section. The performance metrics consist of Matthews correlation coefficient (MCC) [37], precision-recall area under curve (PR AUC) and receiver operating characteristic area under curve (ROC AUC), which is best used to determine the accuracy of imbalanced dataset because of its robustness [38] and accuracy and F-measure that are obtained from the confusion matrix used to determine how well a classification has performed [39] by reporting the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN).

The data presented in Table 2 exhibits the details of the average performance of the classification based on F-measure, MCC, AOC RUC, PR AUC and accuracy after the process was repeated five times coupled with changing of random seed values from 1 to 5.

A recent research [8] on the likelihood of diabetes prediction dataset was compared with the present investigation (Table 3). The results of accuracy, MCC, *F*-Measure, ROC AUC and PR AUC criteria for the models were obtained according to the tenfold cross-validation and percentage split methods compared to the previous study.

Based on the results obtained, we have been able to show not only that better accuracy can be obtained from handling imbalanced dataset but also, more accurate result could be obtained through feature selection. Our results also show that the random forest algorithm performs better than the other algorithms.

This aspect is encouraging for healthcare industry as this model is not expensive and time-consuming to use. As it does not require the use of lab reagents and technical skills unlike OGTT and HbA1c test, therefore, this model can also be put to use to detect diabetes most especially at its early stage in remote areas where health facilities are not accessible.

Method	*F*-measure	MCC	AOC RUC	PR AUC	Accuracy
Random forest (cross-validation)	0.983	0.9654	0.999	0.999	98.3055
Naïve Bayes (cross-validation)	0.8954	0.789	0.9564	0.9564	89.58324
KNN (cross-validation)	0.9792	0.9582	0.9866	0.9836	97.91668
J48 (cross-validation)	0.9674	0.9344	0.9688	0.9582	96.75006
Random forest (split method)	0.979	0.9582	0.9992	0.9992	97.91668
Naïve Bayes (split method)	0.916	0.835	0.9702	0.9698	91.66666
KNN (split method)	0.979	0.9584	0.9874	0.9836	97.91668
J48 (split method)	0.9552	0.9112	0.9576	0.9452	95.55556

Table 2. Summary of performance measure of the classification using cross-validation and percentage split (80:20) methods

Reference	Method	*F*-measure	MCC	AOC RUC	PR AUC	Accuracy
[8]	Random forest + cross-validation	0.9782	0.9534	0.9986	0.9986	97.80768
Ours	Random forest + cross-validation	0.983	0.9654	0.999	0.999	98.3055
[8]	Naïve Bayes + cross-validation	0.8764	0.7462	0.9472	0.9472	87.53846
Ours	Naïve Bayes + cross-validation	0.8954	0.789	0.9564	0.9564	89.58324
[8]	J48 + cross-validation	0.9552	0.906	0.9656	0.9528	95.49998
Ours	J48 + cross-validation	0.9674	0.9344	0.9688	0.9582	96.75006
[8]	Random forest + percentage split	0.9748	0.9474	0.9968	0.9968	97.5
Ours	Random forest + percentage split	0.979	0.9582	0.9992	0.9992	97.91668
[8]	Naïve Bayes + percentage split	0.8908	0.772	0.9554	0.9566	89.03846
Ours	Naïve Bayes + percentage split	0.916	0.835	0.9702	0.9698	91.66666
[8]	J48 percentage split	0.948	0.8914	0.9568	0.9428	94.8077
Ours	J48 + percentage split	0.9552	0.9112	0.9576	0.9452	95.55556
[11]	KNN	0.9621	–	–	–	95.19
Ours	KNN	0.979	0.9584	0.9874	0.9836	97.91668
[11]	Random forest	0.95	–	–	–	97.0
Ours	Random forest	0.979	0.9582	0.9992	0.9992	97.91668

Table 3. The performed studies on detection of diabetes at early stage dataset

Also, all the 13 features work hand in hand for the detection of diabetes as these individual symptoms can be traced to some other diseases as well.

Polyuria and polydipsia attributes are important because the kidney is also affected when people have diabetes. In cases such as fever, diarrhea when the patient is thirsty once water has been drunk, the thirst will be eliminated. But for diabetes this is not so because high blood sugar mounts pressure on the kidneys. The kidneys in turn produce more urine to limit the excess sugar; therefore, causing dehydration, which of course leads to thirst. The cycle continues frequently until the kidneys are weakened and unable to function properly [14, 40]. Sudden weight loss, occasioned by insufficient insulin is one of the early symptoms of diabetes. It (insufficient insulin) prevents the body from producing glucose from the bloodstream to be used as energy in the body's cells. So, when insulin is insufficient in the body, it leads to burning of fat and muscle for energy, which reduces the total weight of the body. Similarly, this in turn leads to polyphagia because when the body lacks enough glucose, it feels more and more hungry. Partial paresis comes in when the body is not able to control the sugar in the blood, which can damage the blood vessels and nerves [14]. The results confirm that age is significant in diabetes diagnosis, which was already declared by another research on this subject matter [14]. Also, gender is crucial in diagnosing diabetes as already stated by a study that had examined carefully men and women separately [41] and observed that it was as a result of higher amount of visceral fat in men. Some important features such as obesity, BMI play no role in some models like the one considered in this study and likewise one (of the models) proposed by [11]. Similarly, a previous study [42] acknowledged that they are not statistically significant with diabetes.

Conclusion

With the rate at which diabetes is increasing among the people, there is a need of detection of diabetes at its early stage. The present study shows the importance of machine learning in the healthcare industry in decision-making and also in reducing the cost of diagnosis. The main contribution of the work is providing a new optimal model for predicting diabetes in its early stage and has also emphasized the importance of feature selection and proper handling of data. It would be interesting – in the future research to know whether body size, height and BMI could be included in the dataset and find the role these parameters play in the detection of diabetes.

The models created with WEKA are available for the readers (https://github.com/OladosuO) for future research purpose.

References

1. Alam TM, Iqbal MA, Ali Y, Wahab A, Ijaz S, Baig TI, Hussain A, Malik MA, Raza MM, Ibrar S and Abbas Z. A model for early prediction of diabetes. Info Med Unlock. 2019; 16. doi: 10.1016/j.imu.2019.100204.
2. Skyler JS, Bakris GL, Bonifacio E, Darsow T, Eckel RH, Groop L, Groop PH, Handelsman Y, Insel RA, Mathieu C, McElvaine AT, Palmer JP, Pugliese A, Schatz DA, Sosenko JM, Wilding JP, Ratner RE. Differentiation of diabetes by pathophysiology, natural history, and prognosis. Diabetes. 2017; 66. doi: 10.2337/db16-0806.
3. Tao Z, Shi A and Zhao J. Epidemiological perspectives of diabetes. Cell Bio Biophys. 2015; 73. doi: 10.1007/s12013-015-0598-4.
4. Alberti KG, Zimmet PZ. Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: diagnosis and classification of diabetes mellitus provisional report of a WHO consultation. Diabet Med, 1998; 15(7): 539-553. doi: 10.1002/(SICI)1096-9136(199807)15:7<539::AID-DIA668>3.0.CO;2-S.
5. WHO, [cited 2020 Dec 18. Online], Available from: www.who.int/health-topics/diabetes#tab=tab_1.

6. Kumari M, Vohra R, Arora A. Prediction of diabetes using bayesian Network. Int J Comput Sci Inf Technol, 2014; 5: 5174-5178.
7. American Diabetes Association website, [cited 2020 Dec 18, [Online]. Available from: http://www.diabetes.org/diabetes-basics/symptoms/.
8. Islam MMF, Ferdousi R, Rahman S, Bushra HY. Likelihood prediction of diabetes at early stage using data mining techniques In: Gupta M, Konar D, Bhattacharyya S, Biswas S (eds), Computer vision and machine intelligence in medical image analysis. Advances in intelligent systems and computing, 2020; 992. doi: 10.1007/978-981-13-8798-2_12.
9. WHO, [cited 2021 May 05 [Online], Available from: www.who.int/news-room/fact-sheets/detail/diabetes.
10. International Diabetes Federation [cited 2020 Dec 21 Online]. Available from: https://www.idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html.
11. Le TM, Vo TM, Pham TN, Dao SVT. A novel wrapper-based feature selection for early diabetes prediction enhanced with a metaheuristic. IEEE Eng Med Bio Soc Sec. 2021; 9. doi: 10.1109/ACCESS.2020.3047942.
12. Diabetes, World Health Organization (WHO), Oct 30, 2018. [cited 2020 Dec 25[Onine]. Available from: https://www.who.int/news-room/ fact-sheets/detail/diabetes.
13. Metzger BE, Lowe LP, Dyer AR, Trimble ER, Chaovarindr U, Coustan DR, Hadden DR, McCance DR, Hod M, Mclntyre HD, Oats JJ, Persson B, Rogers MS, Sacks DA. Hyperglycemia and adverse pregnancy outcomes. N Engl J Med. 2008; 358. doi: 10.1056/NEJMoa0707943.
14. Sneha N, Gangil T. Analysis of diabetes mellitus for early prediction using optimal features selection. J Big Data. 2019; 6. doi: 10.1186/s40537-019-0175-6.
15. Tigga NP and Garg S, Prediction of type 2 diabetes using machine learning classification methods. International Conference on Computational Intelligence and Data Science (ICCIDS 2019), Procedia Computer Science. 2020; 167: 706-16. doi: 10.1016/j.procs.2020.03.336.
16. Ramachandran A. Know the signs and symptoms of diabetes. Indian J Med Res. 2014; 140: 579-81.
17. Bishop CM. Pattern recognition and machine learning (information science and statistics). New York: Springer-Verlag. ISBN 978-0-387-31073-2, 2006.
18. Awan SE, Bennamoun M, Sohel F, Sanfilippo FM, Dwivedi G. Machine learning-based prediction of heart failure readmission or death: implications of choosing the right model and the right metrics. ESC Heart Fail. 2019; 6(2): 428-35. doi: 10.1002/ehf2.12419.
19. Oladimeji OO, Oladimeji O. Predicting survival of heart failure patients using classification algorithms. JITCE (J Info Tech Comp Eng). 2020; 04(02). doi: 10.25077/jitce.4.02.90-94.2020.
20. Joloudari JH, Saadatfar H, Dehzangi A, Shamshirband S. Computer-aided decision-making for predicting liver disease using PSO-based optimized SVM with feature selection. Info Med Unlock; 17. 2017. doi:10.1016/j.imu.2019.100255.
21. Ahmad G, Khan MA, Abbas S, Athar A, Khan BS, Aslam MS. Automated diagnosis of hepatitis B using multilayer mamdani fuzzy inference system. J Health Eng. 2019; 2019: 13-18. doi: 10.1155/2019/6361318.
22. Wang Y, Kung L, Gupta S, Ozdemir S. Leveraging big data analytics to improve quality of care in healthcare organizations: a configurational perspective. Br J Manag. 2019; 30: 362-88. doi: 10.1111/1467-8551.12332.
23. Domingos P. A few useful things to know about machine learning. Commun ACM. 2012; 55(10): 78-87. doi: 10.1145/2347736.2347755.
24. Yang M, Nataliani Y. A feature-reduction fuzzy clustering algorithm based on feature-weighted entropy. IEEE Trans Fuzzy Syst. 2017; 26(2): 817-35. doi: 10.1109/tfuzz.2017.2692203.
25. Chen R, Sun N, Chen X, Yang M, Wu Q. Supervised feature selection with a stratified feature weighting method. IEEE Access. 2018; 6: 15087-98. doi: 10.1109/ACCESS.2018.2815606.

26. Imani, M, Ghassemian H. Feature extraction using weighted training samples. Geosci Rem Sens Lett IEEE. 2015; 12: 1387-91. doi: 10.1109/lgrs.2015.2402167.

27. Cai J, Luo J, Wang S, Yang S. Feature selection in machine learning: a new perspective. Neurocomputing. 2018; 300: 70-79. doi: 10.1016/j.neucom.2017.11.0.77.

28. Mojrian S, Pinter G, Joloudari JH, Felde I, Nabipour N, Nadai L, Mosavi A. Hybrid machine learning model of extreme learning machine radial basis function for breast cancer detection and diagnosis; A multilayer fuzzy expert system. 2020 RIVF International Conference on Computing and Communication Technologies (RIVF), 2020; 1-7. doi: 10.1109/RIVF48685.2020.9140744.

29. Liu S, Yao J, Zhou C, Motani M. Feature selection based on unique relevant information for health data. 2018 IEEE International Conference on Bioinformatics and Biomedicine (BBIM), 2018; 687-692. doi: 10.1109/BBIM.2018.8621163.

30. Liu H, Motoda H, SURI: Feature extraction, construction and selection: a data mining perspective. New York: Springer Science-Business Media; 1998.

31. UCI. Machine learning repository, [cited 2020 Nov 17[Online] Available from: https://archive.ics.uci.edu/ml/index.php.

32. Han J, Pei J, Kamber M. Data mining: concepts and techniques. 3rd ed. Morgan Kaufmann: Elsevier; 2012. doi: 10.1016/C2009-0-61819-5.

33. Larose DT, Larose CD. Discovering knowledge in data: an Introduction to data mining. Hoboken: John Wiley and Sons; 2014. doi: 10.1002/9781118874059.

34. Hu G, Xi T, Mohammed F, Miao H. Classification of wine quality with imbalanced data. 2016 IEEE International Conference on Industrial Technology (ICIT). 2016; 1712-1217. doi: 10.1109/ICIT.2016.7475021.

35. Crisóstomo J, Matafome P, Santos-Silva D, Gomes AL, Gomes M, Patricio M, Letra L, Sarmento-Ribeiro AB, Santos L, Seica R. Hyperresistinemia and metabolic dysregulation: a risky crosstalk in obese breast cancer. Endocrine. 2016; 53(2): 433-42. doi: 10.1007/s12020-016-0893-x.

36. WEKA, [cited 2020 Nov 18 [Online] Available from: www.cs.waikato.ac.nz/ml/weka (accessed 03 April 2020).

37. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Bio Bio Acta (BBA)-Pro Struct, 1975; 405(2): 442–51. doi:10.1016/0005-2797(75)90109-9.

38. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PloS One. 2015; 10(3). doi: 10.1371/journal.pone.0118432.

39. Diez P. Smart wheelchairs and brain-computer interfaces. Cambridge, MA: Elsevier; 2018, ISBN: 978-0-12-812892-3.

40. Health Online, [cited 2021 Feb 8. Online] Available from: https://www.healthonline.com/health/diabetes/3-ps-of-diabetes.

41. Nordstrom A, Hadrevi J, Olsson T, Franks PW, Nordstrom P. Higher prevalence of type 2 diabetes in men than in women is associated with differences in visceral fat. J Clinic Endo Meta, 2016; 101: 3740-46. doi:10.1210/jc.2016-1915.

42. Li Y, Li H, Yao H. Analysis and study of diabetes follow-up data using a data mining-based approach in new urban area of unrunqi Xinjiang, China, 2016-2017. Comp Math Methods Med, 2018; 2018. doi: 10.1155/2018/7207151.

Corresponding author

Oladosu Oyebisi Oladimeji can be contacted at: oladimejioladosu@gmail.com