Assessment of some selected Machine Learning Performance Metrics in the Prediction of type 2 Diabetes

*[1]Julius O. Ogunniyi, [2] Justice O. Emuoyibofarhe, [3]John B. Oladosu and [4]Micheal M. Olamoyegun
[1,3]Department of Computer Engineering, Elizade University, Ilara-Mokin, Nigeria
[2]*Department of Information Systems, LAUTECH, Ogbomoso, Oyo State, Nigeria*
[4]*Department of Internal Medicine, LAUTECH, Ogbomoso, Oyo State, Nigeria*

julius.ogunniyi@elizadeuniversity.edu.ng, eojustice@gmail.com,**dryemi@yahoo.com**, jbjohn@lautech.edu.ng

Received: 17-AUGUST-2024; Reviewed: 20-SEPT-2024; Accepted: 29-SEPT-2024
https://dx.doi.org/10.4314/fuoyejet.v9i3.1

ORIGINAL RESEARCH

Abstract— The performance of machine learning models is crucial in the healthcare domain, as high-performing models ensure accurate diagnostics, effective treatments, and improved patient outcomes thereby enhancing overall healthcare quality. However, researchers often face uncertainty in selecting the appropriate metrics to evaluate predictive models. Therefore, this research aimed to assess selected performance evaluation metrics used in machine learning applications across three different datasets. This study utilized datasets from three sources and three machine-learning algorithms. Logistic regression (LR), naïve Bayes (NB), and CATBoost (CATB) were the classification algorithms used in this work. With accuracy, the area under the curve (AUC), recall, precision, F1-score, kappa, and the Matthews correlation coefficient (MCC) as metrics, the system was constructed using the Python programming language. The accuracy, AUC, Recall, Precision, F1-Score, Kappa, and MCC of LR, NB, and CATB were 78.27%, 0.7529, 0.1484, 0.5433, 0.2210, 0.1426 and 0.1871; 83.42%, 0.8998, 0.8989, 0.8455, 0.8659, 0.6482 and 0.6656; and 97.57%, 0.9741, 0.9789, 0.9798, 0.9789, 0.9503 and 0.9516, respectively on dataset 3. The study evaluated the effectiveness of commonly used machine learning metrics in predicting type 2 diabetes, highlighting the risks of relying solely on accuracy for model evaluation. The study's findings can help machine learning engineers choose the right assessment metric for a given task.

Keywords— Accuracy, AUC, Machine learning, Performance metrics, Type 2 diabetes.

1 INTRODUCTION

Diabetes is a disease that occurs when the body either does not produce enough insulin or cannot effectively use the insulin it produces (Ajani *et al*., 2020). There are four main types of diabetes: Type 1, Type 2, Gestational, and double diabetes (Olamoyegun *et al*., 2020). Among these, Type 2 Diabetes (T2D) is the most prevalent and can lead to serious complications if not addressed early. With the increasing global prevalence of T2D, it presents a significant challenge to the healthcare system (El-Kebbi *et al*., 2021).

Early and accurate prediction of T2D is critical to enhancing preventive measures and treatment plans, ultimately improving patient outcomes (Alanazi, 2022). In recent years, machine learning (ML) models have emerged as powerful tools in predictive healthcare analytics, offering advanced capabilities to identify patterns and predict disease onset (Ibrahim and Saber, 2023). However, the effectiveness of these models largely depends on the performance metrics used to evaluate them. Selecting appropriate metrics is crucial for ensuring the predictive power, reliability, and overall quality of these models.

*Corresponding Author: julius.ogunniyi@elizadeuniversity.edu.ng

Section B- ELECTRICAL/COMPUTER ENGINEERING & COMPUTING SCIENCES
Can be cited as:
Ogunniyi O. J., Emuoyibofarhe O. J, Oladosu B. J, Olamoyegun M. M, (2024). Assessment of some selected Machine Learning Performance Metrics in the Prediction of type 2 Diabetes. FUOYE Journal of Engineering and Technology (FUOYE**JET), 9(3), 452-457.** https://dx.doi.org/10.4314/fuoyejet.v9i3.1

This research focuses on the assessment of some selected ML performance metrics to evaluate their effectiveness in predicting T2D. Various metrics, including accuracy, precision, recall, F1 score, and AUC, were examined to offer a thorough assessment of the model's functionality. Understanding the strengths and limitations of each metric is essential for researchers and healthcare practitioners to make informed decisions when developing and deploying ML-based models (Zhou *et al*., 2021). The study aims not only to enhance the predictive capabilities of ML models for T2D but also to underscore the critical role of metric selection in the model development process.

In this work, seven (7) performance metrics were examined across three (3) ML models, using three (3) different datasets. The focus of the models is on predicting T2D, one of the deadliest diseases worldwide (Zhou *et al*., 2020). By conducting analysis and comparison of performance metrics, this research aspires to provide valuable insights into the evaluation of predictive models, which can be applied in various healthcare scenarios. Ultimately, this study aims to contribute to the development of more robust predictive models for T2D, thereby promoting better patient care and management.

2 RELATED WORKS

Flach (2019), identified gaps in current evaluation procedures and emphasized the development of a robust ML measurement theory. He proposed a new method that a solid theory of ML metrics could generate, stressing those fundamental characteristics, such as classification

skills, may remain hidden without advanced techniques. Handelman *et al*., (2019) focused on the role of ML and AI in radiology and medicine and aimed to empower clinicians to evaluate ML applications in medical practice more effectively. They identified misunderstandings among medical professionals about ML. The authors provided an overview of ML concepts and evaluation metrics and emphasized the importance of transparency in ML methods and algorithms.

Gong (2021) worked on ML models in various classification problems and performance evaluation metrics. He proposed a new evaluation metric combining results from three performance measures, which he claimed are more consistent than traditional accuracy metrics. The developed framework for model evaluation focused on improved consistency in classification tasks.

Sharma and Shah (2021) reviewed ML techniques applied to diabetes prediction. They discussed various ML methodologies and compared their performance. The ML algorithms used in the work were SVM, LR, DTs and ANN. They highlighted the challenges of data inadequacy and deployment in diabetes prediction. They stressed the future potential of ML methods to enhance diabetes prediction and treatment, especially through deep learning models.

Rady *et al*., (2021) also developed models with eight ML algorithms with 521 records of dataset. They compared the performance of these models and concluded that Random Forest was the best-performing algorithm with a 98% accuracy in outsmarting LR, SVM, RF, DT, Adaptive boosting classifier, KNN, and NB.

Naidu *et al*., (2023) focused on the need for accurate measurement and evaluation of ML classifiers in real-world scenarios. They discussed the strengths and limitations of popular metrics (accuracy, precision, F1, and recall). The work highlighted the need for alternative metrics like AUC and Kappa statistics for a deeper understanding. The authors also advocated for standardized measurement procedures to improve the reliability of ML models in practice.

Shrivastava et al., (2023) reviewed the effectiveness of ML methods in predicting diabetes. They examined key aspects of ML methods which include feature selection, data preprocessing, and evaluation metrics.

Rainio et al., (2024) focused on the evaluation of ML models for researchers with limited statistical knowledge. The authors addressed challenges in evaluating ML models, especially for non-experts, and offered practical solutions for accurate model assessment. The work also offered guidance on model comparison, statistical testing, and metric interpretation.

In summary, Handelman et al. (2019) and Sharma and Shah (2021) highlighted the use of machine learning (ML) in healthcare, specifically in the area of diabetes prediction, emphasizing its practical implementation and therapeutic significance.

In-depth talks on evaluation metrics were given by Flach (2019), Gong (2021), and Naidu *et al*. (2023). Each study suggested enhancements or pointed out drawbacks with the metrics that are currently in use (accuracy, AUC, Kappa statistics, etc).

Specifically focusing on diabetes prediction through machine learning, Sharma and Shah (2021), Rady *et al*. (2021), and Shrivastava et al. (2023) used different techniques for feature selection, model validation, and performance comparison.

New theoretical approaches to evaluation were put forth by Flach (2019) and Gong (2021), while Rainio *et al*. (2024) provided workable solutions for ML model evaluation, especially for people with no statistical background.

While existing research has emphasized the importance of knowledge and appropriate metric usage in ML across various fields, the focus has often been on the models themselves rather than the evaluation metrics. This oversight creates a gap in understanding how different performance metrics affect the assessment of models, especially in the context of predicting T2D. therefore, this study aims to evaluate seven selected ML performance metrics to understand better their impact on the effectiveness of predictive models of T2D.

3 METHODOLOGY

This work aimed to develop predictive models and evaluate them using selected evaluation metrics, to assess the performance of these metrics. An experimental approach was used in this work. Three sets of datasets were generated, models were formulated, model implementation was carried out, and model evaluation was carried out considering the performance metrics. Figure 1 shows a block schematic of the model. Each element in the block diagram is further reported in the subsections that follow.

3.1 DATA COLLECTION

There are three (3) datasets used in this work, which are sourced from three different locations. The first dataset is the diabetes dataset from the Kaggle repository, which is 768 in number. The dataset consists of eight (8) risk factors (independent variables), namely, number of pregnancies (pregnancies), glucose, blood pressure (blood pressure), skin thickness (skin thickness), insulin, body mass index (BMI), diabetes pedigree function (DiabetesPedigreeFunction), and age.

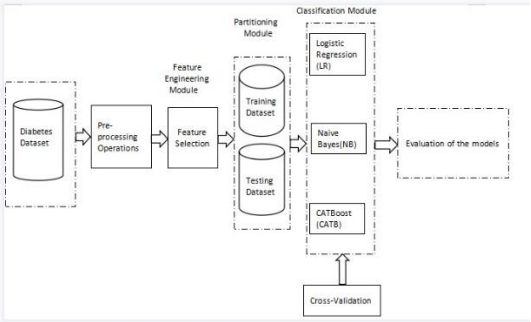

Figure 1: Block diagram of the developed model

http://dx.doi.org/10.46792/fuoyejet.v9i2.C engineering.fuoye.edu.ng/journal

The dependent variable in the dataset is Outcome.

The second dataset was collected from the health records of Ladoke Akintola University of Technology (LAUTECH) Teaching Hospital (8.15150 W, 4.25259E), Mainspring Hospital (8.13333 W, 4.26667E), Ogbomoso, and State General Hospital, Akure (7.10615 W, 4.84665E). There was a total of two hundred and fifty-two (252) records with the following risk factors (independent variables): fasting blood sugar (FBS), BMI, waist circumference, age in years, sex, family history of diabetes, history of excessive urine, regular exercise, history of excessive food and previous history of diabetes. The dependent variable was label class. The third dataset was collected from the Irewolede Community (8.08333 W, 4.18333E) in Ogbomoso during a diabetes awareness program. The data included body mass index (BMI), age, sex, family history of diabetes, hypertensive status, smoking status, alcohol consumption status, regular exercise, waist circumference, and waist-hip ratio. The outcome was the dependent variable. This dataset has a total of one thousand, two hundred and two (1,202) records.

3.2 PREPROCESSING OPERATIONS

In ML solutions, the data preprocessing phase is essential. Selecting risk factors, dealing with missing data, and encoding text and category data, feature selection are all part of it. The risk factors included in this work were identified with the assistance of medical specialists and literature. Depending on the kind of missing value, the mode or mean was used to fill in the missing data. One-hot encoding was employed to encode the categorical features in the datasets. Because the filter method does not overfit the data and has a low computing time, it was used in the feature selection process. A conventional split ratio of 4:1 was used for training and testing datasets, designating 80% of the data for ML algorithm training and 20% for model performance testing.

3.3 MODEL FORMULATION AND TRAINING

The formulation of a predictive model was carried out using the ML algorithms considering the variables of the datasets used. For mathematical models for each of the ML algorithms, the following variables were used to represent the risk factors from each of the datasets.

Dataset 1: Pregnancies=X_1, Glucose= X_2, Blood Pressure= X_3, Skin Thickness= X_4, Insulin= X_5, BMI= X_6, Diabetes Pedigree Function= X_7, Age= X_8 and Outcome= Y

Dataset 2: FBS= X_1, BMI= X_2, waist circumference= X_3, Waist-hip age= X_4, sex= X_5, family history of diabetes= X6, history of excessive urine= X_7, regular exercise= X_8, history of excessive food intake= X_9, previous history of diabetes= X_{10} and class=Y

Dataset 3: Body mass index= X_1, age= X_2, sex= X_3, family history of diabetes= X_4, hypertensive status= X_5, smoking status= X_6, alcohol consumption status= X_7, regular exercise status= X_8, waist circumference= X_9, waist–hip ratio= X_{10} and outcome =Y

1. **Formulation of models using a logistic regression algorithm**

The models were formulated using the sigmoid function in equation (1) and the logistic regression equation in equation (2).

The logistic regression equation is given in equation (1):

$$P(Y = 1 | X_i, .. X_n) = \frac{1}{1 + \exp(-z)} \quad (1)$$

$$z = b_0 + b_i * X_i + .. + b_n * X_n \quad (2)$$

2. **Formulation of models using the Naïve Bayes algorithm**

The models with naïve Bayes algorithms were formulated using Bayes' theorem in equation (3).

$$P(Y|X_i, .., X_n) = \frac{P(Y).X_i|Y \ldots X_n|Y}{P(X_i) \ldots P(X_n)} \quad (3)$$

Were

i=1,2, 3..., n.

n= Number of risk factors in a dataset, X_i=risk factor variable, Y= predicted output, b_0 =bias, b_i=coefficient for inputX_i

3. **Formulation of models using the CATBoost algorithm.**

Equations (4), (5), (6), and (7) were used in the formulation of the models based on the CATBoost algorithm as follows:

a. The model was initialized with a constant value:

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma) \quad (4)$$

b. For m=1 to M:

The pseudo residuals were computed using:

$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \quad for\ i = 1, \ldots, n. \quad (5)$$

The base learner, such as a tree $h_m(x)$ was trained using the training data by fitting the pseudo residuals set$\{(x_i, r_{im})\}_{i=1}^{n}$

By tackling the following one-dimensional optimization issue, a multiplier γ_m was calculated:

$$\gamma_m = -\arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)) \quad (6)$$

c. The model was updated as follows:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \quad (7)$$

d. $F_M(x)$ is the predicted value that was computed.

The hyperparameter tuning was performed using grid search techniques.

In this case, y_i is the target value at i is a multiplier, and $h_m(x)$ is a base learner. L(y, F(x)) is a differentiable loss function.

3.4 IMPLEMENTATION OF THE MODELS

The formulated models were implemented using the Python programming language because of its rich libraries and in-built functions that support the three ML algorithms chosen for this work. The Python program

http://dx.doi.org/10.46792/fuoyejet.v9i2.C engineering.fuoye.edu.ng/journal

was run on an AMD Ryzen 5 2500U with Radeon Vega Mobile Gfx, 2 GHz, 4 Core(s), and 16 GB of RAM.

3.5 EVALUATION METRICS

In this work, the following are the performance evaluation metrics used based on the confusion matrix in Table 1.

a. **Accuracy:** It is the ratio of corre

b. ctly predicted observations to the total observations as given in Equation (8) is the formula for the accuracy.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \quad (8)$$

Table 1: Confusion Matrix

		Actual Values	
		Positive	**Negative**
Predicted values	Positive	TP	FP
	Negative	FN	TN

c. **Precision:** Precision is defined as the ratio of correctly predicted positive observations to all predicted positive observations. The equation contains it (9).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (9)$$

d. **Recall:** Recall is defined as the proportion of correctly anticipated positive observations to all observations made in the actual class. The recall formula is represented by equation (10).

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (10)$$

e. **F1 score:** The precision and recall weighted averages add up to the F1 score. The F1 scoring formula is found in Equation (11).

$$\text{F1 Score} = 2 * \frac{\text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}} \quad (11)$$

f. The **Matthews** correlation coefficient (MCC) determines the classification quality and is divided into two categories. The MCC value offered a binary correlation coefficient between the expected and detected classifications. The MCC formula is given in equation (12).

$$MCC = \frac{TP\ X\ TN - FP\ X\ FN}{(TP + FP)(TP + FN)(TN + FP)(TN + FN)} \quad (12)$$

g. **Kappa Statistic:** Cohen's kappa is another name for the kappa statistic. In reality, it measures a variable's ability to reproduce itself. In equation (13), the formula is given.

$$K = \frac{P_o - P_e}{1 - P_e} = 1 - \frac{1 - P_o}{1 - P_e} \quad (13)$$

Here, P_o= Observed Agreement and P_e= Expected Agreement.

h. **Area under the curve:** Calculating a definite integral between two points yields the area under a curve between those two points. Equation (14) gives the area A under the curve of f from a to b.

$$A = \int_a^b f(x)dx \quad (14)$$

Before releasing the model findings, the 10-fold cross-validation was used to prevent the issue of over-fitting and model bias.

4 RESULTS AND DISCUSSION

4.1 RESULTS

Tables 2, 3, and 4 provide the evaluation results of the constructed models along with the values of all the metrics employed in this research. The Confusion matrix generated from dataset 1 was given in Figures 2a, 2b and 2c. The results of every model on Dataset 1 are compiled in Table 2. Table 4 displays the outcomes of the same models using Dataset 3, while Table 3 displays the models' results using Dataset 2. Section 4.0 of this paper discusses the observations derived from the results.

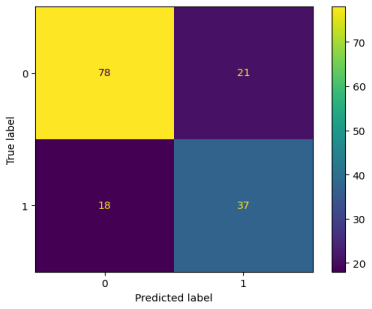

Figure 2a: Confusion matrix of Logistic Regression model on Dataset 1

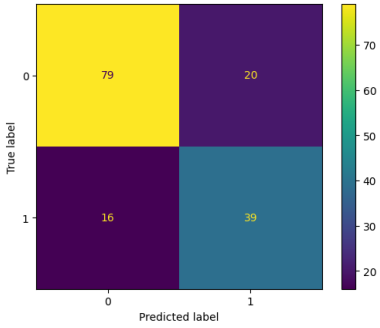

Figure 2b: Confusion matrix of Naïve Bayes model on Dataset 1

http://dx.doi.org/10.46792/fuoyejet.v9i2.C engineering.fuoye.edu.ng/journal

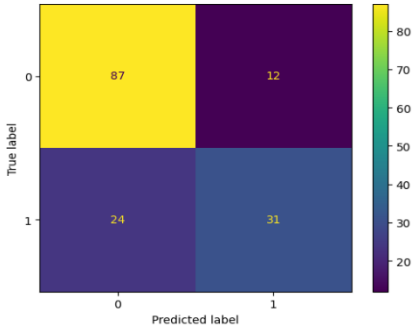

Figure 2c: Confusion matrix of CATBoost model on Dataset 1

Table 2: Summary of the results of all the models with Dataset 1

S/N	Model	Accuracy (%)	AUC	Recall	Precision	F1	Kappa	MCC
1	Logistic Regression	74.68	0.7212	0.6727	0.6379	0.6549	0.4551	0.4555
2	Naïve Bayes	76.62	0.7535	0.7091	0.6610	0.6842	0.4990	0.4998
3	CATBoost	76.62	0.7303	0.5636	0.7209	0.6327	0.4650	0.4725

Table 3: Summary of the results of all the models with Dataset 2.

S/N	Model	Accuracy (%)	AUC	Recall	Precision	F1	Kappa	MCC
1	Logistic Regression	77.01	0.7254	0.0939	0.4417	0.1532	0.0831	0.1167
2	Naïve Bayes	73.05	0.6724	0.4971	0.4208	0.4490	0.2751	0.2802
3	CATBoost	90.60	0.9032	0.6591	0.9073	0.7622	0.7054	0.7203

Table 4: Summary of the results of all the models for Dataset 3

S/N	Model	Accuracy (%)	AUC	Recall	Precision	F1	Kappa	MCC
1	Logistic Regression	78.27	0.7529	0.1484	0.5433	0.2210	0.1426	0.1871
2	Naïve Bayes	83.42	0.8998	0.8989	0.8455	0.8659	0.6482	0.6656
3	CATBoost	97.57	0.9865	0.9789	0.9798	0.9789	0.9503	0.9516

4.2 DISCUSSION

From the results presented in section 4.1, on average, the accuracy metric has the highest value, revealing the models' performance on all the datasets used. But on dataset 1, Naïve Bayes and CATBoost models have the same accuracy of 76.62%, making it difficult to determine the best model between the two. Still, the value of AUC, recall F1, Kappa and MCC revealed that Naïve Bayes outperformed CATBoost on Dataset 1. The results obtained on Datasets 2 and 3 were similar to the ones obtained on Dataset 1. Some authors (Wu *et al.*, 2018; Rady *et al*, 2021; Joshi and Chandra, 2021; Olusanya *et al*., 2022) based their model performance on accuracy metric, the result of this work has revealed the danger of using only accuracy as a performance evaluation metric. This assertion supported the conclusions arrived at by Wang *et al*., (2020) and Ismail *et al*., (2022). The AUC values varied between the models and datasets, with the AUC values being highest in 5 out of the 9 models evaluated, which shows that the AUC is more reliable than the accuracy in judging the performance of predictive models, especially in predicting T2D. This finding corroborates the conclusions drawn in the works of Lotfaliany *et al*., (2019); Battineni, *et al*., (2019); Tigga and Garg, (2020); and Shahriare *et al*., (2020). Kappa statistics consistently revealed the lowest values for the models that were evaluated on the three (3) datasets. These results closely matched MCC findings for every model, indicating that MCC and kappa statistics are both trustworthy measures for evaluating how well models predict type 2 diabetes. The Recall, Precision, and F1 metrics yielded intermediate and reasonable results across all the evaluated models on all the datasets used.

5.0 CONCLUSION

This work has assessed seven (7) metrics on three ML algorithms with three (3) different datasets. The study highlights the limitations of relying solely on accuracy as a performance metric for predictive models while reinforcing the reliability of AUC, Kappa and MCC in the evaluation of ML models, particularly for T2D prediction. From this assessment, it is clear that accuracy alone is insufficient to determine the performance of models. The consistent results from Kappa and MCC confirm their reliability in assessing predictive capabilities. Also, this study emphasizes the importance of using diverse metrics for robust evaluations of predictive models in healthcare.

Future research should explore more diverse datasets, including real-world clinical data, and consider deep learning models as alternative ML algorithms with the performance evaluation metrics used in this work.

REFERENCES

Alanazi, R. (2022). Identification and prediction of chronic diseases using machine learning approach. Journal of Healthcare Engineering, 2022, 2826127. https://doi.org/10.1155/2022/2826127

Ajani, G. O., Gabriel-Alayode, O. E., Atolani, S. A., Soje, M. O., Olamoyegun, M. A., Olarewaju, T. M., & Ajetunmobi, O. A. (2020). Pattern of Dysglycaemia and Family Risk Factors for Diabetes Mellitus among Patients Attending General Outpatient Clinic of Federal Teaching Hospital Ido-Ekiti, Ekiti State, Nigeria. *European Journal of Medical and Health Sciences*, 2(6).

Battineni, G., Sagaro, G. G., Nalini, C., Amenta, F., and Tayebati, S. K. (2019). Comparative Machine-Learning Approach: A Follow-Up Study on Type 2 diabetes predictions by Cross-Validation Methods. *Machines*, 7(4), 74. https://doi.org/10.3390/machines7040074.

El-Kebbi, I. M., Bidikian, N. H., Hneiny, L., and Nasrallah, M. P. (2021). Epidemiology of type 2 diabetes in the Middle East and North Africa: Challenges and call for action. World Journal of Diabetes, 12(9), 1401–1425. https://doi.org/10.4239/wjd.v12.i9.1401

Flach, P. (2019). Performance Evaluation in Machine Learning: The

Good, the Bad, the Ugly,and the Way Forward. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 9808-9814. https://doi.org/10.1609/aaai.v33i01.33019808.

Gong, M., (2021). A Novel Performance Measure for Machine Learning Classification. International Journal of Managing Information Technology (IJMIT), Vol.13, No.1, Available at SSRN: https://ssrn.com/abstract=3807764

Handelman G.S , Hong K. K., Ronil V. C., Amir H.R. , Shiwei H., Mark B., Michael J. L, and Hamed A. (2019).'Peering into the black box of artificial intelligence: Evaluation metrics of machine learning methods', *American Journal of Roentgenology*, 2019, 212(1), pp. 38–43. doi:10.2214/ajr.18.20224.

Ibrahim, M. S., and Saber, S. (2023). Machine Learning and Predictive Analytics: Advancing Disease Prevention in Healthcare. Journal of Contemporary Healthcare Analytics, 7(1), 53–71. Retrieved from https://publications.dlpress.org/index.php/jcha/article/view/16

Ismail, L., Materwala, H., Tayefi, M. *et al.* (2022). Type 2 Diabetes with Artificial Intelligence Machine Learning: Methods and Evaluation. *Arch Computat Methods Eng* **29**, 313–333. https://doi.org/10.1007/s11831-021-09582-x

Joshi, R. D., and Chandra K. D. (2021). "Predicting Type 2 Diabetes Using Logistic Regression and Machine Learning Approaches" *International Journal of Environmental Research and Public Health* 18, No. 14: 7346. https://doi.org/10.3390/ijerph18147346

Lotfaliany, M., Hadaegh, F., Asgari, S., Mansournia, M. A., Azizi, F., Oldenburg, B., and Khalili, D. (2019). Non-nonvasive risk prediction models in identifying undiagnosed type 2 diabetes or predicting future incident cases in the Iranian population. *PubMed*, *22*(3), 116–124. https://pubmed.ncbi.nlm.nih.gov/31029067

Naidu, G., Zuva, T., and Sibanda, E. M. (2023). A Review of Evaluation Metrics in Machine Learning Algorithms. In *Computer Science On-line Conference* 2023, (pp. 15-25). Cham: Springer International Publishing.

Olamoyegun M.A, Ala O.A, and Ugwu E. (2020). Coexistence of type 1 and type 2 diabetes mellitus: a case report of "double" diabetes in a 17-year-old Nigerian girl. Pan Afr Med J [Internet].;37. Available from: http://dx.doi.org/10.11604/pamj.2020.37.35.25191

Olusanya, M. O., Ropo E. O., Meenu G., and Matthew A. A.(2022). "Accuracy of Machine Learning Classification Models for the Prediction of Type 2 Diabetes Mellitus: A Systematic Survey and Meta-Analysis Approach" International Journal of Environmental Research and Public Health 19, No. 21: 14280. https://doi.org/10.3390/ijerph192114280

Rady, M., Moussa, K., Mostafa, M., Elbasry, A., Ezzat, Z., and Medhat, W. (2021). Diabetes Prediction Using Machine Learning: A Comparative Study. In *2021 3rd Novel Intelligent and Leading Emerging Sciences Conference (NILES),* (pp. 279-282). IEEE.

Rainio, O., Teuho, J. and Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. *Sci Rep,* **14**, 6086. https://doi.org/10.1038/s41598-024-56706-x

Sharma, T., and Shah, M. (2021). Machine Learning Techniques for Diabetes Detection: A Comprehensive Review. Visual Computing for Industry, Biomedicine, and Art, 4(30). https://doi.org/10.1186/s42492-021-00097-7

Shahriare S., M., Atik, S.T., and Moni, M.A. (2020). A Novel Hybrid Machine Learning Model to Predict Diabetes Mellitus. In: Uddin,M.S., Bansal, J.C. (eds) Proceedings of International Joint Conference on Computational Intelligence. Algorithms for Intelligent Systems. Springer, Singapore. 2020, https://doi.org/10.1007/978-981-15-3607-6_36

Shrivastava, P., Kumari, A., Kumari, S., and Bajaj, P. (2023). A comprehensive review on the prediction of diabetes disease using machine learning. In Proceedings of the 11th International Conference on Intelligent Systems and Embedded Design (ISED), (pp. 1-6). IEEE. DOI: 10.1109/ISED59382.2023.10444546

Tigga, N. P., and Garg, S. (2020). Prediction of Type 2 Diabetes using Machine Learning Classification Methods. Procedia Computer Science, 167, 706–716. doi:10.1016/j.procs.2020.03

Wang, L., Xiaoya W., Angxuan C., Xian J., and Huilian C. (2020). "Prediction of Type 2 Diabetes Risk and Its Effect Evaluation Based on the XGBoost Model" *Healthcare* 8, No. 3: 247. https://doi.org/10.3390/healthcare8030247

Wu, H., Yang, S., Huang, Z., He, J., and Wang, X. (2018). Type 2 diabetes mellitus prediction model based on data mining. *Informatics in Medicine Unlocked*, *10*, 100–107. https://doi.org/10.1016/j.imu.2017.12.006

Zhou, H., Myrzashova, R. & Zheng, R. Diabetes prediction model based on an enhanced deep neural network. J Wireless Com Network 2020, 148 (2020). https://doi.org/10.1186/s13638-020-01765-7

Zhou, J., Gandomi, A. H., Chen, F., & Holzinger, A. (2021). Evaluating the quality of Machine Learning explanations: A survey on methods and metrics. Electronics, 10(5), 593. https://doi.org/10.3390/electronics10050593