Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

Comparative Analysis of Machine Learning Algorithms for Diabetic Disease Identification

Dhasaradhan Kaveripakam[1,*], Jaichandran Ravichandran[1]

[1] Department of Computer Science and Engineering, Aarupadai Veedu Institute of Technology, Vinayaka Mission's Research Foundation (Deemed to be University), Paiyanoor, Chengalpattu District, Tamil Nadu 603104, India

ARTICLE INFO

Article history:
Received 28 September 2023
Received in revised form 27 March 2024
Accepted 11 April 2024
Available online 12 May 2024

Keywords:

PIMA Indian diabetic dataset; Machine learning algorithms; Jupyter notebook and sci-kit libraries

ABSTRACT

This article introduces a comparative analysis of different types of machine learning algorithms (MLAs) used for diabetic disease identification. Today machine learning algorithms are a major role in solving and identifying the different type of diseases in the medical sector. In the early prediction of diabetic to easily treat physicians and protect from other diseases in patients seven types of MLAs such as support vector machine (SVM), decision tree (DT), logistic regression (LGR), Gradient boost method (GDBM), k-nearest neighbour (KNN), XG boost (XGBM) and random forest (RF) are used for diabetic identification. PIMA Indian diabetic dataset (PIMAIDD) is used to train and test the MLAs. Confusion matrix, accuracy (ACR), precision (PCN), recall (RCL), f1-score (FSC), receiver operating curve (ROC) and K-fold cross-validation are the metrics used for performance evaluation of MLAs and experiments are implemented by Jupyter notebook and python sci-kit libraries. Six types of test cases were conducted whereas test case 4 (70%-30%) was well performed in which RF reported better results in diabetic identification that differentiates from other machine learning metric scores.

1. Introduction

Diabetes is a type of disease that is occurring by increasing the high-level blood sugar in our body. Blood glucose (sugar) is the major energy source of our body and it comes from the food we eat. Insulin is a type of hormone that is produced by Pancreas juice. It extracts glucose from our blood and sends it into the body cells and cells use it as energy. Sometimes when the body doesn't produce enough insulin the glucose is not absorbed by the cells and stays in the bloodstream. As a result, the glucose doesn't reach the cells and the blood contains high-level glucose which is said to be diabetes. There are different types of diabetes such as Gestational diabetes, Type 1 diabetes and Type 2 diabetes [16]. It also leads to diseases such as heart disease, stroke, kidney disease, eye disease, etc. Diabetic patients face symptoms like blurry vision, frequent urination, increased hunger, weight loss, excessive thirst, etc.

* *Corresponding author.*
E-mail address: dhasarath@gmail.com

https://doi.org/10.37934/araset.45.1.4050

Today most of the hospitals maintain the patients' health records in a computerized database, it is easy to train and test the MLAs to identify and diagnose the diseases. In this literature, various MLAs such as SVM, LGR, GDBM, RF, KNN, XGBM and DT are used for diabetic identification. True Negatives (TN), False Positives (FP), True Positives (TP), False Negatives (FN), ACR, PCN, RCL, FSC, ROC and K-fold cross-validation is used for metrics performance evaluation of MLAs. Using PIMAIDD, MLAs can be Trained and tested.

In this paper, we have trained and tested various MLAs such as SVM, LGR, GDBM, RF, DT, KNN, and XGBM in Diabetic identification. PIMAIDD collected from the UCI machine learning repository consists of 768 records of data. After data pre-processing 8 rows of outlier data are removed and the remaining 760 records of data are used for training and testing the MLAs which contain 263 diabetic patient data and 497 non-diabetic person data. There are 9 attributes were use in the research such as Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age and Outcome [14]. MLAs are implemented by Python, pandas, sci-kit learn, numpy, etc., libraries and using Jupyter notebook. MLAs experiments are conducted in six types of test cases, ROC-AUC [5] score and tenfold cross-validation. The performance of MLAs is evaluated using TP, TN, FN, FP, ACR, PCN, RCL, FSC, ROC and K-fold cross-validation as metrics [16-18,21]. We found test case 4 with 30% testing data and 70% training data performed well, in this test case result RF is better for diabetic identification compared to other MLAs in all six test cases.

2. Literature Review

Haram Kim *et al.,* [1] developed automatic diabetic disease prediction application using machine learning algorithms. RTML diabetic disease dataset used train and test MLAs, precision, recall, accuracy, f-score, and ROC AUC score used metrics to evaluate MLAs. Based on the metric results XGBM perform better results with other MLAs.

Raja Krishnamoorthi *et al.,* [2,11] developed MLAs such as DT, SVM and RF algorithms for diabetic disease prediction. PIMAIDD is used for training and testing machine algorithms. The author proposed a unique intelligent diabetes mellitus prediction framework (IDMPF) for diabetic disease prediction. Performance of MLAs ACR, PCN, RCL, FSC and ROC used as metrics. The author proposed an IDMPF algorithm that performs better accuracy compared to other algorithms.

For diabetic disease prediction, Nazin Ahmed *et al.,* built various MLAs such as RF, SVM, DT, GDBM, NB, LGR, and K-NN in diabetic disease prediction. Pima diabetic dataset is used to train and test the models. Efficiency and effective the machine learning algorithm's accuracy is used as a metric. SVM is high accuracy compared with other MLAs [3].

Within the dataset Jobeda Jamal Khanam and Simon Y. Foo developed machine diabetic disease prediction using three MLAs such as LGR, SVM and Artificial Neural Network (ANN), author implemented these algorithms in the Weka tool, Jupyter notebook and Python sci-kit libraries [4]. Performance evaluation MLAs metrics used confusion matrix, ACR, PCN, RCL, FSC and k-fold cross-validation. ANN obtained high accuracy of 88.57% differentiated from other MLAs.

Leila Ismail *et al.,* [5] implemented various MLAs NB, GDBM, RF, SVM, LGR, DT and K-NN using Python and Jupyter notebook for diabetic type 2 prediction. The author trained and tested MLAs using three types of datasets namely PIMA, UCI and MIMIC III . The performance analysis of the MLAs metrics used ACR, PCN, RCL and FSC. RF and LGR performed the better metric score to differentiate from other MLAs.

Within the dataset, Victor Chang *et al.,* [6] developed an Internet of Medical Things (IoTM) using MLAs J48, RF and NB for diabetic disease prediction. PIMAIDD was used to train and test the MLAs. Efficiency and effectiveness of MLAs metrics used true positive, true negative, false positive, false

negative, ACR, PCN, RCL, FSC, ROC and AUC. RF performs better accuracy in 3-factor and NB performs better accuracy, ROC and AUC score in 5-factor.

For diabetic type 2 disease prediction [7] Aishwarya Mujumdara and Dr. Vaidehi V used MLAs such as DT, RF, Gaussian NB, LDA, SVC, Extra Trees, AdaBoost, Perceptron, LGR, GDBM, Bagging and KNN and PIMAIDD train and test these algorithms. Metrics ACR, PCN, RCL and FSC used evaluation of MLAs. LGR obtained better accuracy of 96% compared with other MLAs. The Pipelining results show an accuracy of 98.1% performance by using the algorithms GDBM, AdaBoost Classifier and RF.

In the work of Roshi Saxena *et al.,* [8,19,22] a system predicting diabetic disease designed by the author used four types of classification MLAs namely Multilayer perception (MLP), KNN, RF and DT predicting diabetic disease. PIMAIDD was used to train and test ML algorithms. Sensitivity, specificity and accuracy were used as evaluation metrics of ML algorithms. The system predicts diabetics based on the metrics results in which RF performs better accuracy compared with other ML algorithms.

Arwatki Chen Lyngdoh *et al.,* [9] developed diabetic disease prediction using five supervised ML algorithms namely KNN, DT, SVM, RF and NB. PIMA Indian diabetic dataset used trained and tested the ML algorithms and evaluation of algorithms accuracy and cross-validation used as metrics. The author examines the model's fitting and underfitting cases RF performs better ACR, PCN, RCL and FSC results than other models KNN, DT, SVM and NB.

3. Methodology

Figure 1 shows a flow chart of the methodology for diabetic disease identification, by conducting six types of test cases using machine learning algorithms with help of Jupyter notebook and Python sci-kit libraries.

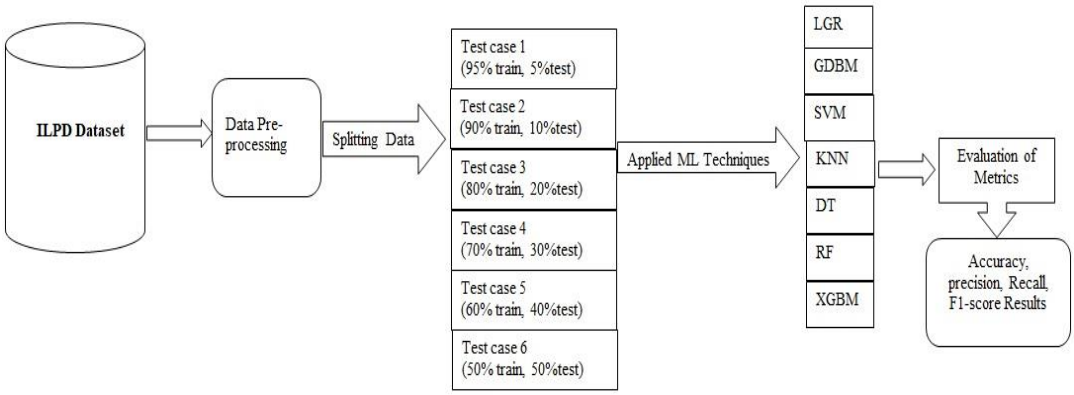

Fig. 1 Methodology flow chart for Diabetic disease

Using metrics ACR, PCN, RCL, FSC, K-Fold Cross-validation results and Roc the Machine Learning algorithms are evaluated in identifying diabetic disease [15].

The model correctly identifies diabetic patient's records in outcomes are true positive (TP), the model correctly identifies non-diabetic patient's records in outcomes are true negative (TN), the model that identifies diabetic disease patient's records incorrectly when the result is true are false positive (FP) and the model that identifies non-diabetic patients records incorrectly are false negative (FN).

Accuracy means the ratio between correctly identifying diabetic disease patients and the total number of diabetic disease patients. Precision means the ratio between models identified diabetic patients and the total number of positive diabetic disease data. The recall is calculated as the ratio between the number of Positive diabetic diseases correctly classified and the total number of Positive

diabetic disease data and false negative data. The F1 score is calculated by dividing precision by recalling harmonically.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

$$\text{F1 Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

A true positive rate (TPR) refers to the ratio between correctly classified positive data and the total number of positive data and the ratio between correctly classified negative data and the number of negative data.

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{6}$$

Tenfold cross-validation means the PIMA dataset is randomly split into ten parts. 1 part of the data was used for testing and 9 parts of the data for training; repeated this technique every time preserving a changed tenth for test data. K-fold cross-validation is a technique for performance evaluation of machine learning algorithms, the dataset is split trained, and tested into the k-number of the fold and calculates the mean of k-folds. In our experiment, k-10 folds are used to train and test the machine learning algorithms. PIDD dataset after pre-processing consists of 768 records of data; 692 records of data are used to train the models and 76 records of data are used to test the models.

ROC (receiver operating characteristic curve) is plotting a curve as the x and y-axis. The x-axis is denoted as FPR and the y-axis denotes TPR. It is a tool used for the performance of MLAs for all threshold values.

4. Experimental Results

MLAs are developed in Jupyter notebook and Python sci-kit libraries. Experiments are conducted in the following test cases as the results are given below.

4.1 Test Case 1

5% testing data and 95% training data which includes 722 records of data training and 38 records of data testing. Training data includes 249 diabetic patient data and 473 non-diabetic data. Testing data contains 14 diabetic patient data and 24 non-diabetic data. Figure 2 shows LGR obtained 92 %

ACR, 91 % PCN, 92 % RCL and 92 % FSC, GDBM obtained 89% ACR, 90% PCN, 89% RCL and 89% FSC, SVM obtained 89% ACR, 89% PCN, 88% RCL and 89% FSC, KNN obtained 87 % ACR, 87% PCN, 87% RCL and 87% FSC, DT obtained 82% ACR, 83% PCN, 82% RCL and 82% FSC, RF obtained 90% ACR, 89% PCN, 90% RCL and 90% FSC, XGBM obtained 89% ACR, 89% PCN, 89% RCL and 89% FSC. Figure 2 shows LGR and RF reported high ACR, PCN, RCL and FSC compared to other MLAs. Results show LGR is better in diabetic identification compared to other MLAs in test case1.

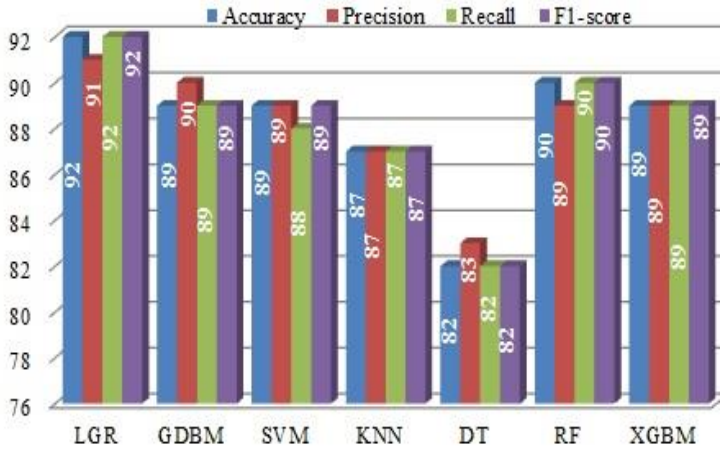

Fig. 2. Results of test case1- Diabetic disease

4.2 Test Case 2

10% testing data and 90% training data which includes 684 records of data in training and 76 records of data in testing. Train data includes 236 diabetic patient data and 448 non-diabetic data. The test dataset includes 27 diabetic disease patients' data and 49 non-diabetic disease data. Figure 3 shows LGR obtained 89 % ACR, 90 % PCN, 89 % RCL and 90 % FSC, GDBM obtained 93% ACR, 93% PCN, 93% RCL and 93% FSC, SVM obtained 88% ACR, 88% PCN, 88% RCL and 88% FSC, KNN obtained 86 % ACR, 85% PCN, 86% RCL and 86% FSC, DT obtained 87% ACR, 88% PCN, 87% RCL and 87% FSC, RF obtained 90% ACR, 90% PCN, 89% RCL and 90% FSC, XGBM obtained 89% ACR, 89% PCN, 89% RCL and 89% FSC. Above Figure 3 GDBM and RF reported high ACR, PCN, RCL and FSC compared to other models. Results show GDBM is better in diabetic identification compared to other MLAs in test case2.

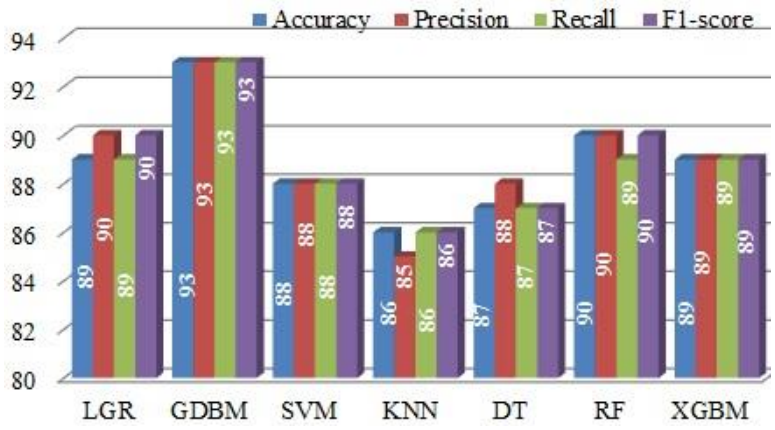

Fig. 3. Results of test case1- Diabetic disease

4.3 Test Case 3

20% test data and 80% train data which includes 608 records of data training and 152 records of data testing. Training data contains 209 diabetic disease patients' data and 399 non-diabetic data. The testing dataset contains 54 diabetic disease patient data and 98 non-diabetic patient data. Figure 4 shows LGR obtained 88 % ACR, 88 % PCN, 88 % RCL and 88 % FSC, GDBM obtained 92% ACR, 92% PCN, 92% RCL and 92% FSC, SVM obtained 87% ACR, 87% PCN, 87% RCL and 87% FSC, KNN obtained 86 % ACR, 86% PCN, 86% RCL and 86% FSC, DT obtained 87% ACR, 82% PCN, 82% RCL and 82% FSC, RF obtained 90% ACR, 90% PCN, 90% RCL and 90% FSC, XGBM obtained 89% ACR, 89% PCN, 89% RCL and 89% FSC. Figure 4 shows GDBM and RF reported high ACR, PCN, RCL and FSC compared to other models. Results show GDBM is better in diabetic identification compared to other MLAs in test case 3**.**

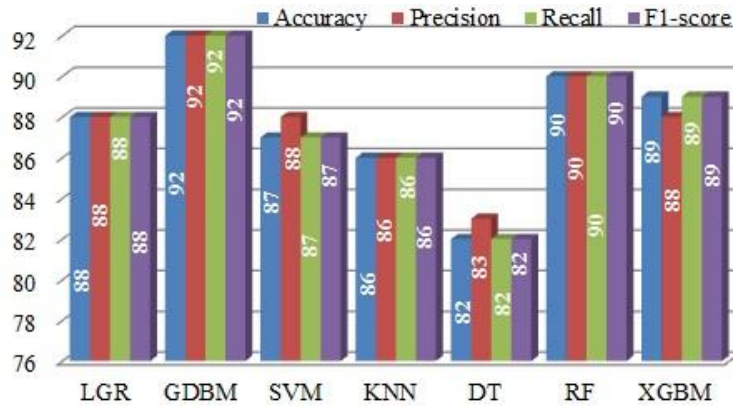

Fig. 4. Results of test case3- Diabetic disease

4.4 Test Case 4

30% test data and 70% train data which includes 532 records of data for training and 228 records of data for testing. Train data contains 182 diabetic patient data and 350 non-diabetic patient data. The test dataset contains 81 diabetic patient data and 147 non-diabetic patient data. Figure 5 shows LGR obtained 88 % ACR, 88 % PCN, 88 % RCL and 88 % FSC, GDBM obtained 89% ACR, 89% PCN, 90% RCL and 89% FSC, SVM obtained 84% ACR, 83% PCN, 84% RCL and 83% FSC, KNN obtained 83 % ACR, 83% PCN, 82% RCL and 83% FSC, DT obtained 83% ACR, 84% PCN, 83% RCL and 82% FSC, RF obtained 93% ACR, 92% PCN, 92% RCL and 93% FSC, XGBM obtained 88% ACR, 86% PCN, 87% RCL and 88% FSC. Figure 5 shows RF and GDBM reported high ACR, PCN, RCL and FSC compared to other models. Results show RF is better in diabetic identification compared to other MLAs in test case 4**.**

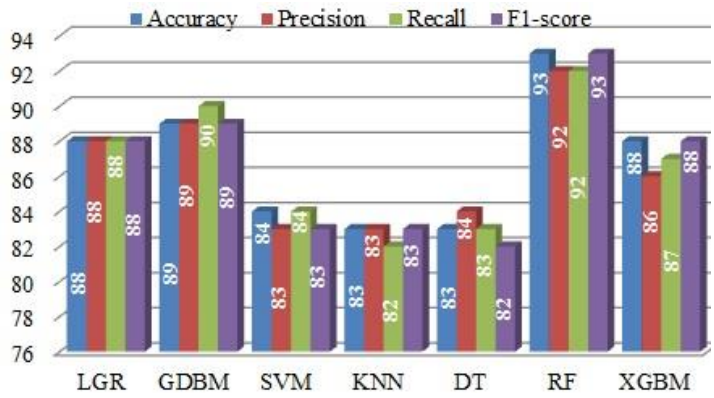

Fig. 5. Results of test case4- Diabetic disease

4.5 Test Case 5

40% testing data and 60% training data which includes 456 records of data for training and 304 records of data for testing. Train data contains 149 diabetic disease patients' data and 307 non-diabetic data. The test dataset contains 114 diabetic patient data and 190 non-diabetic data. Figure 6 LGR obtained 87 % ACR, 87 % PCN, 86 % RCL and 87 % FSC, GDBM obtained 89% ACR, 88% PCN, 87% RCL and 89% FSC, SVM obtained 84% ACR, 85% PCN, 85% RCL and 84% FSC, KNN obtained 83 % ACR, 82% PCN, 81% RCL and 81% FSC, DT obtained 87% ACR, 87% PCN, 87% RCL and 87% FSC, RF obtained 90% ACR, 90% PCN, 89% RCL and 90% FSC, XGBM obtained 88% ACR, 87% PCN, 88% RCL and 88% FSC. Figure 6 shows RF and GDBM reported high ACR, PCN, RCL and FSC compared with other models. Results show RF is better in diabetic identification compared to other MLAs in test case 5.

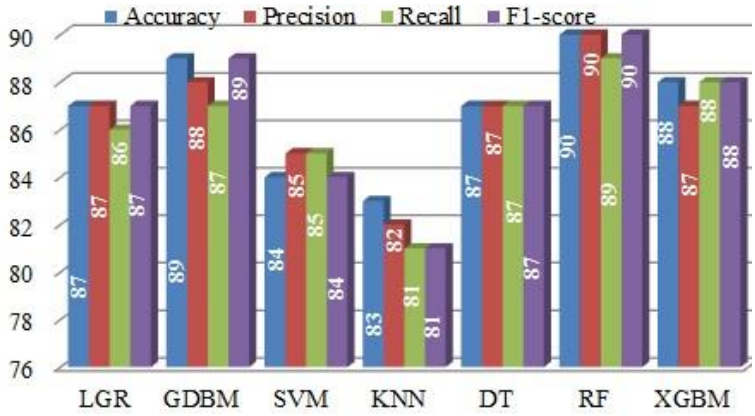

Fig. 6. Results of test case 5- Diabetic disease

4.6 Test Case 6

50% testing data and 50% training data which includes 380 records of data for training and 380 records of data for testing. Training data contains 128 diabetic patient data and 252 non-diabetic data. The dataset contains 135 diabetic patient data and 245 non-diabetic data. Figure 7 shows LGR obtained 86 % ACR, 85 % PCN, 85 % RCL and 86 % FSC, GDBM obtained 89% ACR, 88% PCN, 89% RCL and 89% FSC, SVM obtained 84% ACR, 84% PCN, 84% RCL and 82% FSC, KNN obtained 82 % ACR, 83% PCN, 83% RCL and 83% FSC, DT obtained 82% ACR, 83% PCN, 83% RCL and 82% FSC, RF obtained 90% ACR, 90% PCN, 90% RCL and 90% FSC, XGBM obtained 88% ACR, 88% PCN, 88% RCL and 88%

FSC. Figure 7 shows RF and GDBM reported high ACR, PCN, RCL and FSC compared with other models. Results show RF is better in diabetic identification compared to other MLAs in test case 6.

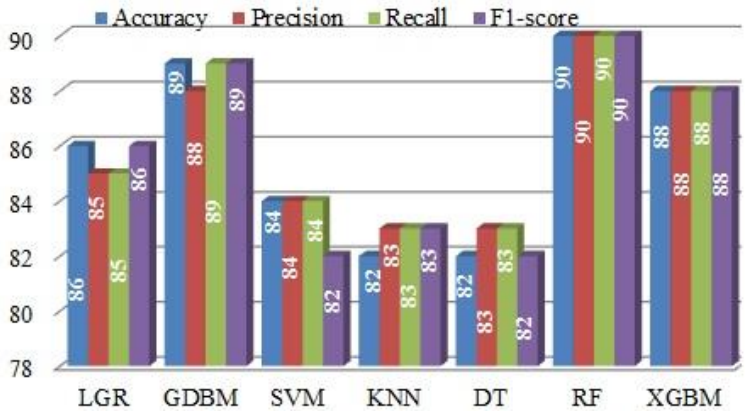

Fig. 7. Results of test case 6- Diabetic disease

Table 1
Six type's test case results

Test Cases	Metrics	Machine learning Algorithms						
		LGR	GDBM	SVM	KNN	DT	RF	XGBM
Case 1	Accuracy	92	89	89	87	82	90	89
	Precision	91	90	89	87	83	89	89
	Recall	92	89	88	87	82	90	89
	F1-Score	92	89	89	87	82	90	89
Case 2	Accuracy	89	93	88	86	87	90	89
	Precision	90	93	88	85	88	90	89
	Recall	89	93	88	86	87	89	89
	F1-Score	90	93	88	86	87	90	89
Case 3	Accuracy	88	92	87	86	82	90	89
	Precision	88	92	88	86	83	90	89
	Recall	88	92	87	86	82	90	89
	F1-Score	88	92	87	86	82	90	89
Case 4	Accuracy	88	89	84	83	83	92	88
	Precision	88	89	83	83	84	92	86
	Recall	88	90	84	82	83	92	87
	F1-Score	88	89	83	83	82	93	88
Case 5	Accuracy	87	89	84	83	87	90	88
	Precision	87	88	85	82	87	90	87
	Recall	86	87	85	81	87	89	88
	F1-Score	87	89	84	81	87	90	88
Case 6	Accuracy	86	89	84	82	82	90	88
	Precision	85	88	84	83	83	90	88
	Recall	85	89	84	83	83	90	88
	F1-Score	86	89	82	83	82	90	88

A ROC is a curve which is plotting points between true positive (TPR) and false positive rate (FPR) with their different threshold values, it is used to perform evaluating machine learning algorithms [6]. From Figure 8 we obtained ROC-AUC score LGR reported 87%, DT reported 86%, SVM reported 81%, KNN reported 81%, XGBM reported 87%, RF reported 92% and GDBM reported 88%. Based on ROC results RF is well-identified whether the person is suffering from diabetes or not.

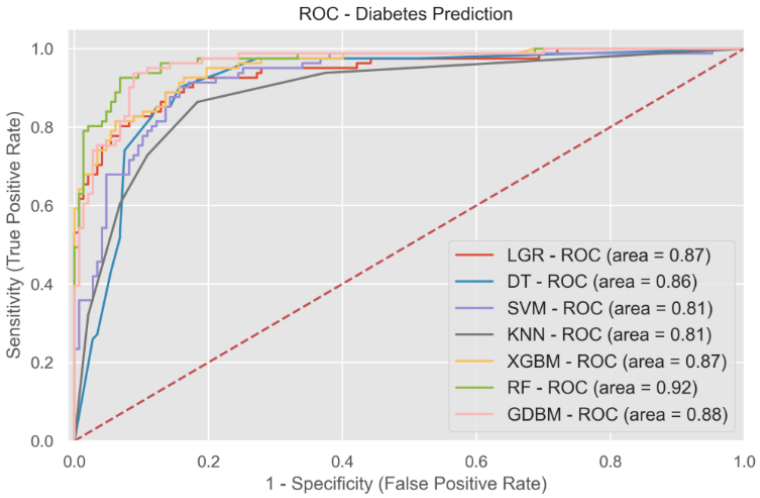

Fig. 8. ROC Curve Results- Diabetic disease

In tenfold cross-validation, the PIMA dataset is divided into ten parts of data [8,13]. One part data is testing and 9 parts of the data are used for training. Finally, calculating the mean value of ten parts of the dataset results in this accuracy results in the performance evaluation of the machine learning model. In Figure 9 tenfold cross-validation means value accuracy LGR obtained 80%, GDBM obtained 88%, DT obtained 70%, KNN obtained 80%, XGBM obtained 88%, SVM obtained 85% and RF obtained 93%. Based on the tenfold cross-validation results RF model performed well in identifying whether the patient is suffering from diabetes or not.

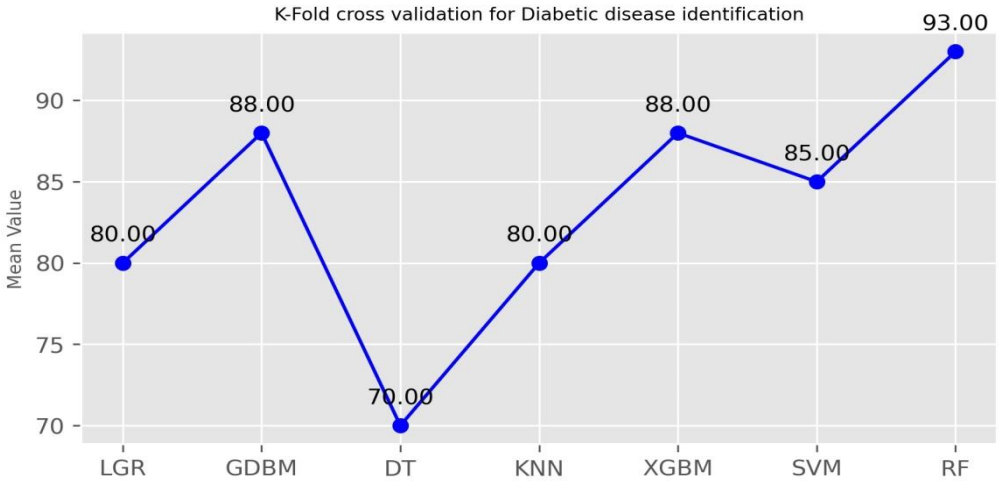

Fig. 9. Tenfold cross-validation results

5. Conclusion

Machine learning algorithms are a major role in the medical sector to detect and identify diseases. Today hospitals maintain their patient's health records in e-data format, so it is easy to test and train the MLAs. In this research, we applied different types of MLAs such as LGR, XGBM, SVM, KNN, DT, XGBM and RF for identifying diabetic disease [10,12]. ACR, PCN, RCL and FSC these metrics used performance and evaluation of machine learning algorithms [20,21]. In this work estimating the algorithms we used different techniques namely tenfold cross-validation, ROC-AUC score and different test case results. According to the experimental results of six types of test case results, test case 4 is 70% training data and 30% performed better accuracy (92%) metric score results. ROC-AUC

(92%) scores and 10 k-fold cross validation (94%) result we compared the experimental findings RF model is better for identifying diabetics.

Acknowledgement
This research was not funded by any grant.

References
[1] Kim, Haram, and Dongsoo Kim. "Deep-Learning-Based Strawberry Leaf Pest Classification for Sustainable Smart Farms." *Sustainability* 15, no. 10 (2023): 7931. https://doi.org/10.3390/su15107931
[2] Krishnamoorthi, Raja, Shubham Joshi, Hatim Z. Almarzouki, Piyush Kumar Shukla, Ali Rizwan, C. Kalpana, and Basant Tiwari. "A novel diabetes healthcare disease prediction framework using machine learning techniques." *Journal of healthcare engineering* 2022 (2022). https://doi.org/10.1155/2022/1684017
[3] Ahmed, Nazin, Rayhan Ahammed, Md Manowarul Islam, Md Ashraf Uddin, Arnisha Akhter, Md Alamin Talukder, and Bikash Kumar Paul. "Machine learning based diabetes prediction and development of smart web application." *International Journal of Cognitive Computing in Engineering* 2 (2021): 229-241. https://doi.org/10.1016/j.ijcce.2021.12.001
[4] Khanam, Jobeda Jamal, and Simon Y. Foo. "A comparison of machine learning algorithms for diabetes prediction." *Ict Express* 7, no. 4 (2021): 432-439. https://doi.org/10.1016/j.icte.2021.02.004
[5] Ismail, Leila, Huned Materwala, Maryam Tayefi, Phuong Ngo, and Achim P. Karduck. "Type 2 diabetes with artificial intelligence machine learning: methods and evaluation." *Archives of Computational Methods in Engineering* (2022): 1-21. https://doi.org/10.1007/s11831-021-09582-x
[6] Chang, Victor, Jozeene Bailey, Qianwen Ariel Xu, and Zhili Sun. "Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms." *Neural Computing and Applications* 35, no. 22 (2023): 16157-16173. https://doi.org/10.1007/s00521-022-07049-z
[7] Mujumdar, Aishwarya, and V. Vaidehi. "Diabetes prediction using machine learning algorithms." *Procedia Computer Science* 165 (2019): 292-299. https://doi.org/10.1016/j.procs.2020.01.047
[8] Saxena, Roshi, Sanjay Kumar Sharma, Manali Gupta, and G. C. Sampada. "A novel approach for feature selection and classification of diabetes mellitus: Machine learning methods." *Computational Intelligence and Neuroscience* 2022 (2022). https://doi.org/10.1155/2022/3820360
[9] Lyngdoh, Arwatki Chen, Nurul Amin Choudhury, and Soumen Moulik. "Diabetes disease prediction using machine learning algorithms." In *2020 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES)*, pp. 517-521. IEEE, 2021. https://doi.org/10.1109/IECBES48179.2021.9398759
[10] Butt, Umair Muneer, Sukumar Letchmunan, Mubashir Ali, Fadratul Hafinaz Hassan, Anees Baqir, and Hafiz Husnain Raza Sherazi. "Machine learning based diabetes classification and prediction for healthcare applications." *Journal of healthcare engineering* 2021 (2021). https://doi.org/10.1155/2021/9930985
[11] Bavkar, Vandana C., and Arundhati A. Shinde. "Machine learning algorithms for diabetes prediction and neural network method for blood glucose measurement." *Indian J Sci Technol* 14, no. 10 (2021): 869-880. https://doi.org/10.17485/IJST/v14i10.2187
[12] Ljubic, Branimir, Ameen Abdel Hai, Marija Stanojevic, Wilson Diaz, Daniel Polimac, Martin Pavlovski, and Zoran Obradovic. "Predicting complications of diabetes mellitus using advanced machine learning algorithms." *Journal of the American Medical Informatics Association* 27, no. 9 (2020): 1343-1351. https://doi.org/10.1093/jamia/ocaa120
[13] Chaki, Jyotismita, S. Thillai Ganesh, S. K. Cidham, and S. Ananda Theertan. "Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review." *Journal of King Saud University-Computer and Information Sciences* 34, no. 6 (2022): 3204-3225. https://doi.org/10.1016/j.jksuci.2020.06.013
[14] Li, Xiaohua, Jusheng Zhang, and Fatemeh Safara. "Improving the accuracy of diabetes diagnosis applications through a hybrid feature selection algorithm." *Neural processing letters* 55, no. 1 (2023): 153-169. https://doi.org/10.1007/s11063-021-10491-0
[15] Ayon, Safial Islam, and Md Milon Islam. "Diabetes prediction: a deep learning approach." *International Journal of Information Engineering and Electronic Business* 13, no. 2 (2019): 21. https://doi.org/10.5815/ijieeb.2019.02.03
[16] Ramadhan, Nur Ghaniaviyanto, and Ade Romadhony. "Preprocessing handling to enhance detection of type 2 diabetes mellitus based on random forest." *International Journal of Advanced Computer Science and Applications* 12, no. 7 (2021). https://doi.org/10.14569/IJACSA.2021.0120726

[17] Fregoso-Aparicio, Luis, Julieta Noguez, Luis Montesinos, and José A. García-García. "Machine learning and deep learning predictive models for type 2 diabetes: a systematic review." *Diabetology & metabolic syndrome* 13, no. 1 (2021): 1-22. https://doi.org/10.1186/s13098-021-00767-9

[18] Basha, C. Bagath, and K. Somasundaram. "A comparative study of twitter sentiment analysis using machine learning algorithms in big data." *International Journal of Recent Technology and Engineering* 8, no. 1 (2019): 591-599.

[19] Kidam, Kamarizan, Siti Aishah Rashid, Jafri Mohd Rohani, Hafizah Mahmud, Hamidah Kamarden, Fateha Abdul Razak, Nurul Nasuha Mohd Nor, and Nur Kamilah Abdul Jalil. "Development of Instrument to Measure the Impact of COVID-19 And Movement Control Order to Safety and Health Competent Person and Training Provider." *Journal of Advanced Research in Technology and Innovation Management* 2, no. 1 (2022): 22-28.

[20] Ramadhan, Nur Ghaniaviyanto, and Ade Romadhony. "Preprocessing handling to enhance detection of type 2 diabetes mellitus based on random forest." *International Journal of Advanced Computer Science and Applications* 12, no. 7 (2021). https://doi.org/10.14569/IJACSA.2021.0120726

[21] Basha, C. Bagath, and K. Somasundaram. "A comparative study of twitter sentiment analysis using machine learning algorithms in big data." *International Journal of Recent Technology and Engineering* 8, no. 1 (2019): 591-599.

[22] Kidam, Kamarizan, Siti Aishah Rashid, Jafri Mohd Rohani, Hafizah Mahmud, Hamidah Kamarden, Fateha Abdul Razak, Nurul Nasuha Mohd Nor, and Nur Kamilah Abdul Jalil. "Development of Instrument to Measure the Impact of COVID-19 And Movement Control Order to Safety and Health Competent Person and Training Provider." *Journal of Advanced Research in Technology and Innovation Management* 2, no. 1 (2022): 22-28.