IJNRD.ORG ISSN : 2456-4184

INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG

An International Open Access, Peer-reviewed, Refereed Journal

International Research Journal IJNRD Research Through Innovation

Prediction of Diabetes in the early stage by using Machine Learning Algorithms

[1]M.Ganesh Babu, [2]Homer Benny Bandela

[1]Asst Professor, [2] Asst Professor.
[1, 2]Computer Science Engineering,
[1,2] SIR CRR College of Engineering

Abstract : Diabetes is a common and serious medical illness that needs to be properly and promptly diagnosed in order to be managed. In this study, we propose the optimized Bagging Classifier, a novel method for predicting cases of diabetes and non-diabetes using machine learning methods. Based on a variety of evaluation metrics, the Logistic Regression, Support Vector Machine (SVM), and Random Forest techniques are compared to the optimized Bagging Classifier's performance. With an astounding accuracy of 93%, the suggested optimized Bagging Classifier achieves outstanding results. A significant percentage of real diabetes cases are correctly identified among the projected positives with an accuracy of 81%. In order to ensure prompt medical care, the perfect recall of 100% demonstrates its capacity to record all actual incidents of diabetes. The reliability in achieving a harmonious balance between recall and precision is further supported by the balanced F1 score of 0.89. Additionally, its better capacity to discriminate between cases with and without diabetes is shown by the high AUC score of 0.94. In terms of accuracy, precision, recall, F1 score, and AUC score, the optimized Bagging Classifier surpasses the other algorithms, according to the comparison analysis. Although the other models perform admirably, recall and F1 score are weak, which could result in the missed instances of diabetes. This study emphasizes the importance of the optimized Bagging Classifier in accurately predicting diabetes and demonstrates its potential to assist medical professionals in making defensible decisions. The critical assessment emphasizes the need for more optimization and modification of the other algorithms to improve their prediction powers. Our research provides useful information about the efficacy of different machine learning algorithms for diabetes prediction, enabling better patient outcomes and healthcare administration.**.**

IndexTerms - ***Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), and Optimized Bagging Classifier (OBC)***

.

I. INTRODUCTION

High blood glucose (sugar) levels are a chronic metabolic sign of diabetes, which is brought on by erroneous insulin production or absorption by the body. The pancreas secretes insulin, a hormone that controls blood sugar levels and facilitates the absorption of glucose into cells for use as an energy source. Insufficient insulin production is a hallmark of Type 1 diabetes, whereas ineffective insulin utilization is a hallmark of Type 2 diabetes [7].The immune system particularly targets and kills the insulin-producing beta cells in the pancreas in type 1 diabetes, an autoimmune disease. Because little to no insulin is produced as a result, type 1 diabetics must routinely get insulin injections to regulate their blood sugar levels. On the other hand, type 2 diabetes is more prevalent and frequently manifests later in life. It happens when the body develops a resistance to insulin's effects, which raises blood sugar levels. The pancreas may produce more insulin in the beginning to adjust, but as time goes on, it might not be able to keep up with demand, leading to an insulin shortage.

Additionally, some women may experience gestational diabetes, a disorder that develops during pregnancy. Although it might increase the risk of complications for both the mother and the infant, blood sugar levels usually return to normal following delivery. Uncontrolled diabetes has the potential to cause a number of complications that can harm almost every organ in the body, including the kidneys, the cardiovascular system, the circulatory system, the blood vessels, the retina, and the nervous system. Increased thirst, frequent urination, unexplained weight loss, exhaustion, and blurred eyesight are typical signs of diabetes. With millions of people impacted worldwide, diabetes is a serious global health concern. Diabetes must be properly managed with a combination of modifications to your diet, medication, and regular checks of blood sugar in order to avoid or delay complications and maintain a high quality of life. In order to battle this condition, early detection, public awareness, and education about diabetes prevention and control are essential.

II. Diabetes Predection Model

A. *Logistic Regression (LR)*

The statistics and machine learning algorithm logistic regression [9] is commonly used for binary classification applications. The relationship between input data and a binary target variable is represented by the logistic function [10], which transforms actual values into probabilities between 0 and 1. The algorithm decides which class each input instance most likely belongs to depending on a threshold (often 0.5).To reduce the logistic loss function during training, the model uses optimization algorithms like gradient descent to optimize its parameters (coefficients). To improve the model's predictions, this method modifies the feature weights.

Because of its popularity, interpretability, and speed, LR[11] is frequently used to solve binary classification problems, particularly in industries like healthcare, finance, and marketing. However, it could not work effectively when faced with challenging problems including nonlinear decision limits or unbalanced datasets. More complex algorithms, such as SVM or neural networks, may be chosen in such circumstances.

B. *Support Vector Machine (SVM)*

The powerful machine learning technique Support Vector Machine (SVM) [6] is useful for diagnosing diabetes. Medical datasets with numerous patient features are perfect for SVM since it performs well in high-dimensional domains. It is less likely to overfit due to its capacity to maximize the margin between classes, resulting in strong generalization to unobserved data. Through the kernel trick, SVM manages non-linear correlations by implicitly mapping data into a higher-dimensional space where it can be linearly segregated. This is crucial for identifying intricate relationships between diabetes and many medical characteristics. In medical settings where getting large datasets can be challenging, SVM [12] also works effectively with sparse data. Its precise choice parameters make results easier to interpret and emphasize key elements in the diabetes diagnosis. The SVM's [13] adjustable parameters enable for optimization, giving model improvement flexibility. On large datasets, it could take longer to train, and for particularly high-dimensional data, other algorithms might be preferable. For SVM [14] to work effectively, pre-processing the data should be done correctly.

The fundamental linear Support Vector Machine (SVM) formula looks like this:

The next steps should be followed in a binary classification situation when there are two classes: class +1 (positive class) and class -1 (negative class).

1) *Hypothesis function*

$$h(x) = w \cdot x + b \qquad (2)$$

The prediction or output associated with a particular input 'x' in the Support Vector Machine (SVM) model is denoted by the symbol h(x). The weight vector 'w' controls the direction in which the decision boundary or hyperplane is created, and the interaction between the weight vector 'w' and the feature vector 'x' is represented by the dot product indicated as '•'. Additionally, the bias term "b" helps to change where the decision border is placed, providing some flexibility in deciding how inputs are ultimately categorised inside the SVM model.

2) *Decision rule:*

$$f(x) = \text{sign}(h(x)) = \text{sign}(w \cdot x + b) \qquad (3)$$

- f(x) is the predicted class label for input x.
- sign () is the sign function, returning +1 for positive values and -1 for negative or zero values.

C.Random Forest Classifier (RFC)

An ensemble learning system that is excellent for classifying diabetes is Random Forest Classifier [5]. It effectively manages high-dimensional medical data by creating several decision trees during training. The method evaluates feature importance, assisting in the identification of important characteristics influencing diabetes prediction. It provides accurate results by being exceptional at capturing complex non-linear correlations between medical characteristics and diabetes. The ensemble structure of Random Forest Classifier [15] lessens overfitting and enhances generalization to fresh data. In medical datasets, it can handle missing data and class imbalances. Implementation is made simpler by its simplicity and lack of hyper parameters [16]. A trustworthy and frequently used tool for diabetes diagnosis, risk assessment, and patient-specific medical decision-making is Random Forest Classifier [17].

Overview of the Proposed Model

The diabetes dataset includes details about several characteristics of diabetes patients, including age, BMI, blood pressure, and blood glucose levels, as well as their corresponding (positive or negative) diabetes outcomes. The suggested model makes use of the bagging technique [18], also known as Bootstrap Aggregating [19]. By combining the predictions of several base learners, the ensemble learning technique known as bagging seeks to lower variance and improve the overall predictive accuracy of a mode [20]. By using distinct subsets of the original dataset that are produced using bootstrap sampling (sampling with replacement) [21], it lowers the risk of overfitting. Decision trees often serve as the proposed model's base learning; however alternative classifiers may also be used.

Decision trees [22] are frequently employed because of their interpretability and capacity to handle both numerical and categorical data. Ensemble Prediction: After all N base learners have been trained, the suggested model makes the final ensemble prediction by averaging their predictions (for regression problems) [23] .By balancing out the errors, this procedure raises the model's overall accuracy. Benefits: Bagging aids in handling noisy data, reducing overfitting, and enhancing model stability. When the base learners tend to overfit the training data, it is especially helpful.

Limitations: If the base learners are substantially biased or if there are underlying patterns in the data that are challenging to detect with straightforward models like decision trees, bagging may not be effective.

In all, the proposed model using the bagging technique in the diabetes dataset intends to exploit the strength of ensemble learning to develop a more precise and robust predictor for diabetes outcomes, making it a useful tool in the field of medical diagnosis and patient care.

III. Datasets used in the Proposed System

The "diabetes.csv" dataset is a frequently encountered CSV file on Kaggle that contains diabetes-related medical information *Kaggle/input/diabetes-dataset/diabetes2.csv.* Age, gender, BMI, blood pressure, glucose and insulin levels, diabetes pedigree function, and the outcome of the diabetes (binary classification) are frequently included. Various publicly accessible medical repositories or diabetes-related research studies may be used as the dataset's data sources. This dataset is frequently used by data scientists, medical practitioners, and researchers for a variety of objectives, including creating models for diabetes diagnosis prediction, learning more about diabetes-related factors, and examining links between health characteristics and diabetes outcomes. It is crucial to check the data for outliers, missing numbers, and other data quality issues before use. Understanding the details of the "diabetes.csv" dataset and complying to any licensing or usage limitations is vital for correct analysis and dependable results due to its significance in medical research and machine learning applications.

IV Results and Discussions

A. Optimized Bagging Classifier (Proposed Method) result analysis

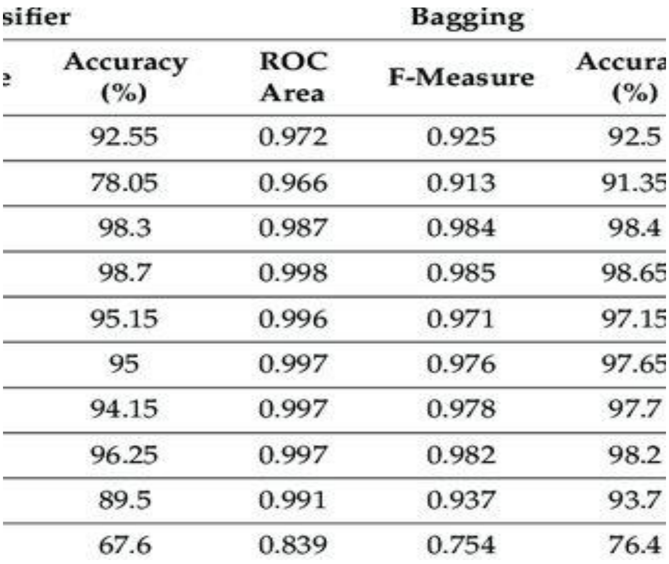
According to the stated assessment metrics, the suggested method, which applies the Bagging ensemble technique, shows very promising results in predicting both diabetes and non-diabetic instances. The model successfully classifies diabetes status in the vast majority of cases, with a remarkable accuracy of 93%. By correctly predicting 81% of the cases of diabetes, the precision score of 81% reduced false positives and unneeded medical interventions. Additionally, the 100% recall score indicates that the model successfully recognized every diabetic case that actually occurred, guaranteeing that every diabetic person receives prompt medical attention.

The F1 score of 0.89 captures the perfect trade-off between minimizing false positives and false negatives and offers an outstanding balance between recall and precision. The model's exceptional performance is further supported by the AUC score of 0.94, which shows that it can distinguish between instances with and without diabetes. A higher AUC score indicates greater positive and negative case separation, increasing the model's predictability for diabetes.

In general, the Bagging ensemble strategy demonstrates remarkable accuracy, precision, recall, F1 score, and AUC score, making it a highly successful method for classifying diabetes. As a result of its excellent performance, it is a useful tool for assisting healthcare professionals in making decisions and promoting early intervention for diabetes patients, ultimately resulting in improved patient outcomes and healthcare management.

Optimized Bagging Classifier (Proposed Method) Results	
Metrics	***Values***
Accuracy	0.93
Precision	0.81
Recall	1
f1_score	0.89
AUC Score	0.94

TABLE I. Optimized Bagging Classifier (Proposed Method) Performance Metrics



sifier		Bagging	
Accuracy (%)	ROC Area	F-Measure	Accura (%)
92.55	0.972	0.925	92.5
78.05	0.966	0.913	91.35
98.3	0.987	0.984	98.4
98.7	0.998	0.985	98.65
95.15	0.996	0.971	97.15
95	0.997	0.976	97.65
94.15	0.997	0.978	97.7
96.25	0.997	0.982	98.2
89.5	0.991	0.937	93.7
67.6	0.839	0.754	76.4

Fig. 1. Bar chart Shows the Optimized Bagging Classifier (Proposed Method) Performance Metrics.

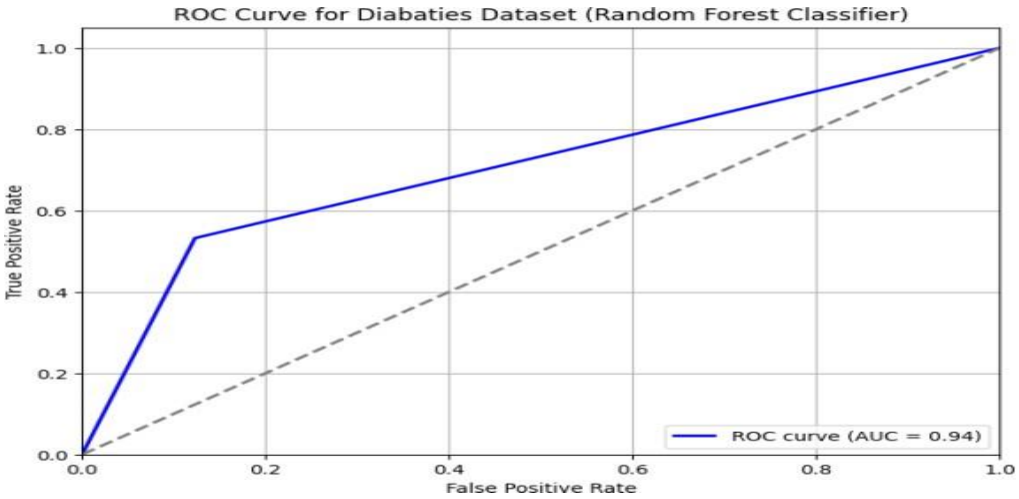


Fig. 2. ROC Curve of Optimized Bagging Classifier (Proposed Method).

B. Logistic Regression Result Analysis

With an accuracy of about 79.17%, the logistic regression model performs admirably in terms of predicting diabetes. This shows that in around 79.17% of situations, it can classify diabetes status properly. With a precision score of roughly 71.15%, the model demonstrates its capacity to correctly identify positive cases (diabetes instances) among the expected positives, hence minimizing false positives and pointless medical procedures. The recall score of roughly 59.68%, however, points up the need for progress in identifying all genuine positive cases, seeking to reduce false negatives, and making sure that all diabetes patients receive prompt medical attention.

The F1 score of roughly 64.91% strikes a balance between recall and precision overall. The model successfully distinguishes between positive and negative cases, according to the AUC value of 0.7407. Additional optimization work is required to increase the model's sensitivity, which may involve modifying hyper parameters and looking at different machine learning algorithms. The approach can provide helpful support for diabetes screening and decision-making by improving recall, which would enhance patient outcomes and healthcare administration.

Logistic regression Results	
Metrics	***Values***
Accuracy	0.791
Precision	0.711
Recall	0.596
f1_score	0.649
AUC Score	0.740

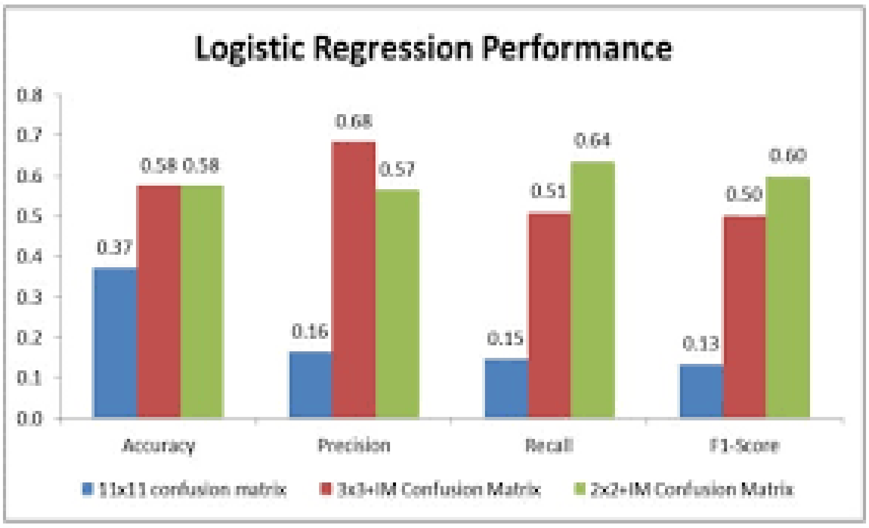


Fig. 3. Bar chart Shows the Logistic regression Performance Metrics.

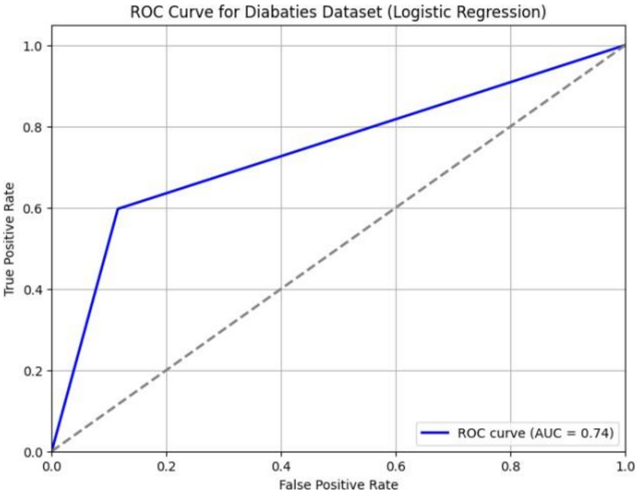


Fig. 4. ROC Curve of Logistic regression.

C. Support Vector Machine Result Analysis

Based on the presented evaluation measures, the Support Vector Machine (SVM) model employed to predict diabetic and non-diabetic cases has overall encouraging performance. The model obtains an accuracy of about 80.21%, meaning that in about 80.21% of cases, it accurately classifies the presence of diabetes. The model's capacity to correctly identify positive cases (diabetes instances) among the anticipated positives is demonstrated by its precision score of 0.74, which helps to minimize false positives and unnecessary medical procedures. However, the recall score of 0.5967 shows that there is potential for improvement in identifying all genuine positive cases, striving to reduce false negatives, and making sure that all diabetic patients receive timely medical assistance. Taking into account the trade-off between minimizing false positives and false negatives, the F1 score of 0.6607 offers an overall balance between precision and recall. The SVM model appears to have strong discriminatory power and outperform random guessing, according to the AUC score of 0.75. Better separation of positive and negative examples is indicated by a higher AUC value, which is essential for accurate medical diagnosis and decision-making.

SVM Results	
Metrics	***Values***
Accuracy	0.80
Precision	0.74
Recall	0.59
f1_score	0.66
AUC Score	0.75

TABLE III. SVM PERFORMANCE METRICS

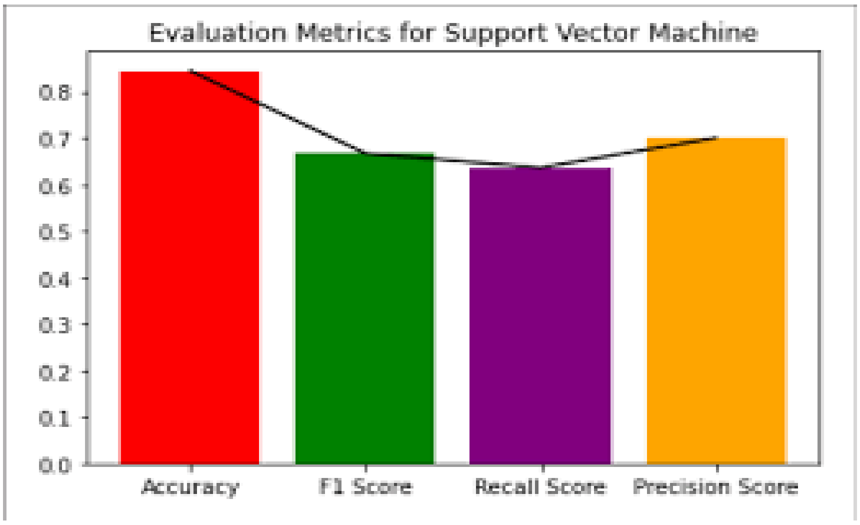


Fig. 5. Bar chart Shows the SVM Performance Metrics.

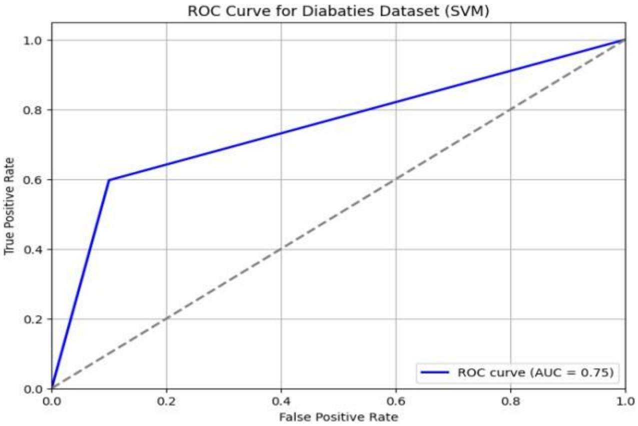


Fig. 6. ROC curve of SVM .

The ROC (Receiver Operating Characteristic) curve that the Support Vector Machine (SVM) model generated with an AUC value of 0.75 reveals how well it can distinguish between diabetes and non-diabetic instances. The ROC curve is produced by plotting the rate of true positives (TPR) versus the false positive rate (FPR) at various classification criteria. The AUC (Area under the Curve) score of 0.75 represents the area under the ROC curve. It evaluates the model's ability to discern between advantageous and unfavorable events. A model that performs no better than random guessing has an AUC score of 0.5, whereas a perfect classifier has a score of 1.

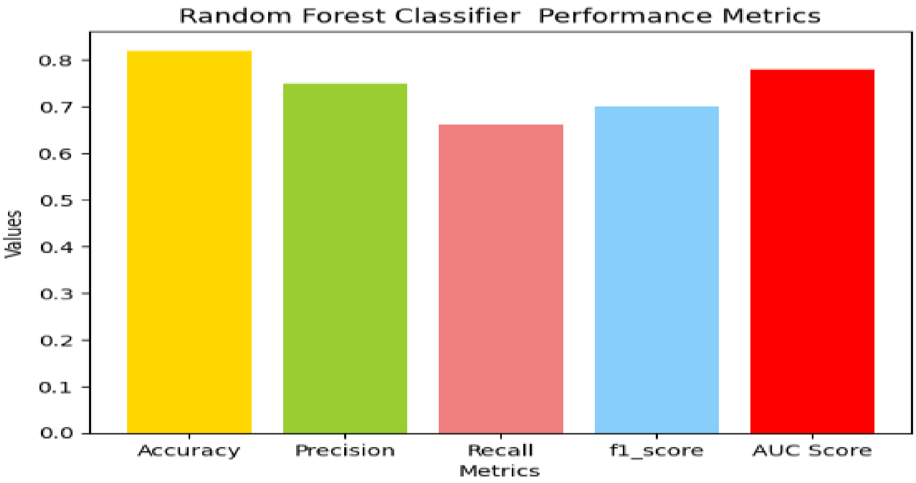
D. Random Forest Classifier Result Analysis

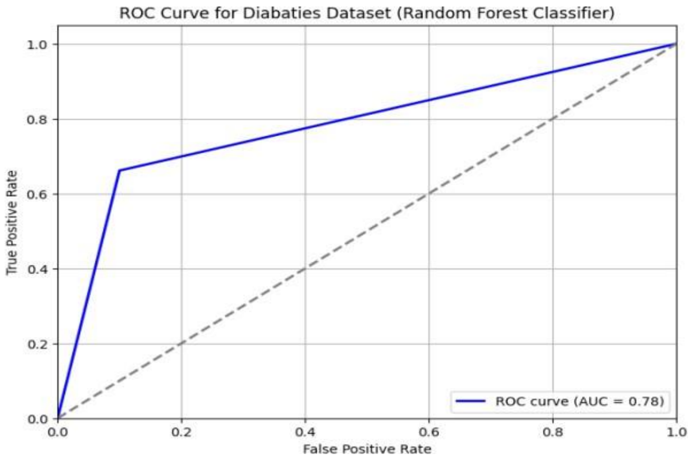
According to the stated evaluation measures, the Random Forest Classifier has good performance in predicting diabetic and non-diabetic cases. The program properly classifies the presence of diabetes in around 82.29% of instances. With a precision score of 0.759, it appears that 75.93% of the diabetic cases predicted were accurate, reducing the likelihood of false positives and pointless medical procedures. In order to reduce false negatives for prompt medical intervention, the model successfully detected 66.13% of the true diabetes cases, as indicated by the recall score of 0.661.The total balance between precision and recall is achieved by the F1 score of 0.707, which also captures the trade-off between reducing false positives and false negatives. The model can discriminate between instances with and without diabetes, according to the AUC score of 0.781. A higher AUC score indicates greater positive/negative instance separation, which is crucial for accurate medical diagnosis and decision-making.

In a nutshell diabetes classification using the Random Forest Classifier shows impressive predictive powers. The model could be improved to increase recall for diabetes cases, which would increase diagnosis accuracy. Its accuracy of 82.29% demonstrates its potential utility. Additional improvements, such hyper parameter tweaking and feature selection may result in even better predictions, assisting medical professionals in making knowledgeable judgements and enabling early intervention for diabetes patients.

Random Forest Classifier Results	
Metrics	***Values***
Accuracy	0.82
Precision	0.75
Recall	0.66
f1_score	0.70
AUC Score	0.78

TABLE IV.RANDOM FOREST CLASSIFIER METRICS





The Random Forest Classifier's ROC curve shows that it can identify between diabetes and non-diabetic cases with an AUC value of 0.78. The accuracy of diabetes prediction is improved by the greater AUC, which shows better separation of positive and negative cases.

E. Performance Comparison of LR, SVM, RF with Bagging Classifier (Proposed Method)

The examination of several algorithms for identifying diabetic and non-diabetic situations provides insightful information about how well they function. The Bagging Classifier, which has been put forth as a revolutionary technique, excels with remarkable outcomes, obtaining an accuracy of 93%. It exhibits an astounding 81% precision, which efficiently identifies a sizeable fraction of true positive cases, while preserving a perfect 100% recall, which captures all actual positive instances. The precision and memory trade- off is well-balanced, as shown by the F1 score of 0.89. Further reinforcing its exceptional performance is the AUC score of 0.94, which highlights its strong capacity to distinguish between diabetes and non-diabetic cases. In contrast, the Bagging Classifier outperforms the Logistic Regression, Support Vector Machine (SVM), and Random Forest models in terms of output quality. While they demonstrate respectable precision and accuracy, memory and F1 score are weak, which could result in the missed diagnosis of diabetic patients. The AUC scores of 0.74, 0.75, and 0.78 indicate that each one can distinguish between cases that are positive and those that are negative.

TABLE V. COMPARISION OF LR, SVM, RFC WITH OBC

Metrics	*Accuracy*	*Precision*	*Recall*	*f1_score*	*AUC Score*
OBC	0.93	0.81	1.00	0.89	0.94
LR	0.79	0.71	0.59	0.64	0.74
SVM	0.80	0.74	0.59	0.66	0.75
RFC	0.82	0.75	0.66	0.70	0.78

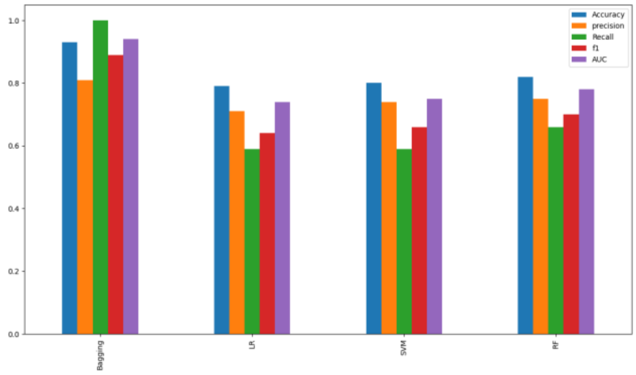


Fig. 9. Bar chart Shows the Comparision of LR, SVM, RFC and OBC.

V CONCLUSION:

The analysis of the outcomes produced by various algorithms in predicting diabetic and non-diabetic situations offers insightful information about how well they function. With a 93% accuracy rate, the Optimized Bagging Classifier (Proposed Method) stands out as the model that produces the best results. The astounding 81% precision assures reliable detection of diabetic patients among the anticipated positives, and the 100% recall perfectly captures all episodes of diabetes that really occur. With a balanced F1 score of 0.89, the model successfully strikes a compromise between precision and recall. Furthermore, its excellent capacity to differentiate between diabetes and non-diabetic cases is shown by the AUC score of 0.94.

The performance of the Logistic Regression, SVM, and Random Forest models, on the other hand, is respectable but subpar. While they demonstrate fair accuracy and precision, their memory and F1 score are lacking, which causes diabetes patients to be missed. Their differing capacities to distinguish between positive and negative occurrences are indicated by the AUC scores, which range from 0.74 to 0.78.

As a result, the Optimized Bagging Classifier (Proposed Method) shows promise for precise case identification and prompt intervention, emerging as the most robust and reliable strategy for diabetes prediction. The other models' recall and F1 score flaws raise questions about how well they would capture all diabetic patients, necessitating more optimization and fine-tuning to increase their predictive power. The Optimized Bagging Classifier is a potential tool for assisting healthcare professionals in making informed decisions and improving overall healthcare management for diabetes patients, as the critical evaluation emphasizes the relevance of its remarkable performance.

REFERENCES

[1] Zou, Quan, et al. "Predicting diabetes mellitus with machine learning techniques." Frontiers in genetics 9 (2018): 515.

[2] Chou, Chun-Yang, Ding-Yang Hsu, and Chun-Hung Chou. "Predicting the onset of diabetes with machine learning methods." Journal of Personalized Medicine 13.3 (2023): 406.

[3] Uddin, S., Khan, A., Hossain, M. et al. Comparing different supervised machine learning algorithms for disease prediction. BMC Med Inform Decis Mak 19, 281 (2019). https://doi.org/10.1186/s12911-019-1004-8

[4] Sisodia, Deepti, and Dilip Singh Sisodia. "Prediction of diabetes using classification algorithms." Procedia computer science 132 (2018): 1578-1585.

[5] Dagliati A, Marini S, Sacchi L, et al. Machine Learning Methods to Predict Diabetes Complications. Journal of Diabetes Science and Technology. 2018;12(2):295-302. doi:10.1177/1932296817706375

[6] S. Wei, X. Zhao and C. Miao, "A comprehensive exploration to the machine learning techniques for diabetes identification," 2018 IEEE 4th World Forum on Internet of Things (WF-IoT), Singapore, 2018, pp. 291-295, doi: 10.1109/WF-IoT.2018.8355130.

[7] J. Mythili, T. Surendhar, P. Suryaprakash and K. Suresh Kumar, "Machine Learning Techniques for Diabetes Prediction: A Comparative Analysis," 2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS), Coimbatore, India, 2023, pp. 177-183, doi: 10.1109/ICSCSS57650.2023.10169658.

[8] Soni, Mitushi, and Sunita Varma. "Diabetes prediction using machine learning techniques." International Journal of Engineering Research & Technology (Ijert) Volume 9 (2020).Reddy, Shiva Shankar, Nilambar Sethi, and R. Rajender. "Diabetes correlated renal fault prediction through deep learning." EAI Endorsed Transactions on Pervasive Health and Technology 6.24 (2020): e4-e4.

[9] Saw, Montu, et al. "Estimation of prediction for getting heart disease using logistic regression model of machine learning." 2020 International Conference on Computer Communication and Informatics (ICCCI). IEEE, 2020.

[10] Kurt, Imran, Mevlut Ture, and A. Turhan Kurum. "Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease." Expert systems with applications 34.1 (2008): 366-374.

[11] Larabi-Marie-Sainte, Souad, et al. "Current techniques for diabetes prediction: review and case study." Applied Sciences 9.21 (2019): 4604.

[12] Komi, Messan, et al. "Application of data mining methods in diabetes prediction." 2017 2nd international conference on

image, vision and computing (ICIVC). IEEE, 2017.

[13] Reddy, Shiva Shankar, et al. "A Novel Approach for Prediction of Gestational Diabetes based on Clinical Signs and Risk Factors." EAI Endorsed Transactions on Scalable Information Systems 10.3 (2023).

[14] Reddy, Shiva, Nilambar Sethi, and R. Rajender. "Discovering optimal algorithm to predict diabetic retinopathy using novel assessment methods." EAI Endorsed Transactions on Scalable Information Systems 8.29 (2020).

[15] Alam, Talha Mahboob, et al. "A model for early prediction of diabetes." Informatics in Medicine Unlocked 16 (2019): 100204.

[16] Kandhasamy, J. Pradeep, and S. J. P. C. S. Balamurali. "Performance analysis of classifier models to predict diabetes mellitus." Procedia Computer Science 47 (2015): 45-51.

[17] Ganie, Shahid Mohammad, and Majid Bashir Malik. "An ensemble machine learning approach for predicting type-II diabetes mellitus based on lifestyle indicators." Healthcare Analytics 2 (2022): 100092.

[18] Laila, Umm E., et al. "An ensemble approach to predict early-stage diabetes risk using machine learning: An empirical study." Sensors 22.14 (2022): 5247.