[1]Amit Dubey/s,

[2] Pritaj Yadav/s

Improving Diabetes Detection Using Machine Learning Random Forest Algorithm

Journal of Electrical Systems

Abstract: - Due to current lifestyle most of the peoples suffering from numerous diseases, heart disease, diabetes, obesity, etc. diabetes is one of the disease is found at an every age group of peoples nowadays. Here discussion is based on diabetes disease detection at an early stage and finding the pattern or recognizing diseases in patients. Artificial intelligence techniques very popular in health care sector, AI based techniques such as machine learning and deep learning techniques are having very tremendous growth in health care and information security sectors for providing best results than traditional techniques. In this paper, we discuss comparative performance analysis between different machine learning techniques among them random forest classifier gives best performance than other techniques.

Keywords: Machine learning, Healthcare, Pattern recognition, Classification, Information security, World Health Organization.

I. INTRODUCTION

Nowadays most of the peoples are suffering from various types of diseases, there are number of reason to increasing the diseases day by day such as junking food, lifestyle, smoking, and some of sue to genetic. In the current scenario there are various techniques and tools are used in healthcare industry to provide goof healthcare services for every patients, now artificial intelligence techniques are very popular for healthcare applications. In currently diabetes disease is very common and affected almost every age group, and gender. According to WHO the number of diabetic patients are increasing day by day, and Considering that diabetes is associated with increased rates of intensive care unit (ICU) admissions and mortality.

Effectively managing diabetes necessitates regular monitoring of blood glucose levels to prevent or delay potential long-term health complications.

Linked to the ailment [3]. The hormone insulin, which is synthesized in the pancreas, has a crucial role in controlling the body's blood glucose levels.

Diabetes has two common types, type 1 diabetes mellitus (T1DM) and type 2 diabetes mellitus (T2DM). Individuals with diabetes for an extended period are at risk of developing several complications such as heart disorders, kidney disease, nerve damage, diabetic retinopathy, and more. Early prediction and intervention can help reduce these risks [1].

The rest of the paper is assembled in Sections. Section II discusses the machine learning techniques, and the role of machine learning techniques in the health care sector to identify diseases in patients. Section III elaborates on the proposed model using a machine learning approach with the proposed model workflow. Section IV discusses the experimental work and comparative result analysis for different machine learning approach with performance parameters study for diabetes disease prediction and finally, Section V consists of the generated conclusion of the shown comparative study and future directions.

II. MACHINE LEARNING IN HEALTHCARE

Machine learning is a sub field of artificial intelligence technique which is used in various applications like healthcare sector, digital image processing, speech and text recognition, face recognition, pattern recognition,

[1]Ph.D. Research Scholar, Department of Computer Science and Engineering, Rabindranath Tagore University Bhopal, M.P.India1

[1]*Corresponding author: amitdubeyppi@gmail.com

[2]Associate Professor, Department of Computer Science and Engineering, Rabindranath Tagore University, Bhopal(M.P.),

security surveillance, and many more. Machine learning is basically divided into some categories like SML, UML, SSML, and RML. Here supervised machine learning (SML) techniques works with known data

and labels, unsupervised machine learning (UML) techniques works with unknown data and here no class of data and labels, semi-supervised machine learning (SSML) techniques works combine the role of SML techniques with UML techniques, and reinforcement machine learning (RML) technique work on the basis of reward.

ML techniques works here for the diabetes disease detection and prediction at an early stage in a patients, in this research work we mention the some supervised machine learning techniques to improve the performance parameters value and enhance the existing model than previous techniques.

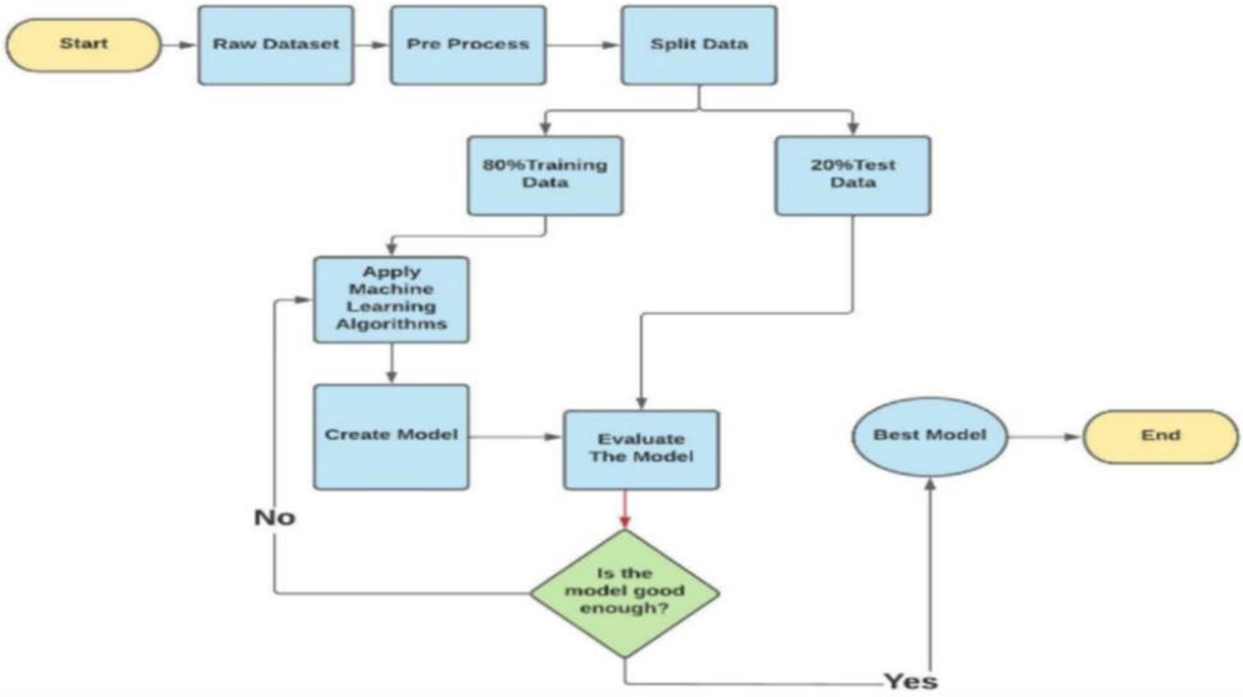


Figure 1:*Diabetes prediction system with ML techniques [8].*

III. PROPOSED WORK

ML techniques are widely used in various applications, healthcare sector or medical imaging is one of them. Healthcare sector is affecting almost every person directly or indirectly in current era, because every person is almost suffering from any types of diseases at any level with any age group, to avoid all the diseases with prevention, and reduce the risk of any disease in life we used ML techniques to provide the results on the basis of past experience and history of patients with increasing accuracy rate and enhancements in performance than existing techniques or approach.

Random forests (RF) employ an ensemble approach during training by constructing numerous individual decision trees. These decision trees collectively contribute to the final prediction, with classification relying on the mode of the classes and regression relying on the mean prediction. This collective decision-making process characterizes them as Ensemble techniques.

In Random Forest, the importance of the features is determined with the help of a matrix and this matrix measures the reduction in impurity of the node,, weighted by the probability of reaching that node. This probability is calculated by considering the number of samples crossing a specific node, divided by the total number of samples. Higher values for this metric indicate greater feature importance.

In the case of Scikit-learn's implementation of Random Forests, it calculates the importance of nodes using Gini Importance, assuming a binary tree structure with only two child nodes.

$$mj_k = X_kE_k - X_{left(k)}E_{left(k)} - X_{right(k)}E_{right(k)-(1)}$$

The importance of node "k," denoted as "mj $_{\text{k}}$," signifies the significance of node "k" in a given context.

“x$_{\text{k}}$” denotes the weighted count of samples visiting node “k”.

“E$_{\text{k}}$” corresponds to the impurity value from node “k”.

"left(k)" specifies the child node that originates from the left split of node "k."

"right(k)" signifies the child node that originates from the right division of node "k."

The procedure for computing the importance of each feature in a decision tree is as follows:

$$fj_{\mathrm{j}} = \frac{\sum \mathrm{k{:}\ node\ k\ splits}\ on\ feature\ j^{nj}k}{\sum \mathrm{k} \in \mathrm{all\ nodes}^{nj}k} --(2)$$

The importance of a feature, represented as "fj $_{\text{j}}$"

The significance of a node, represented as " njK," can be normalized to a range between 0 and 1 by dividing each of these values by the total sum of all feature importance values.

$$norm\ fj_j = \frac{fj_j}{\sum k \in \mathrm{all}\ features^{\ fj_j}} ----(3)$$

In a Random Forest, the collective feature importance is ascertained by computing the mean importance across all individual trees. This is accomplished by adding up the importance values of the feature in each tree and subsequently dividing the sum by the total number of trees.

$$RF\ fj_j = \frac{\sum k \in \mathrm{all}\ trees^{\ normsfj_{jk}}}{T} ----(4)$$

Certainly, here are the statements rewritten without plagiarism:

"RFfjj" signifies the feature importance of feature "j" calculated across all trees within the Random Forest model.

"normsfjjk" represents the normalized feature importance of feature "j" within the context of tree "k."

"T" represents the total count of trees within the Random Forest model.

Within each decision tree in Spark, the importance of a feature is determined by aggregating the gain, which is then adjusted based on the number of samples that traverse through the respective node.

$$fj_j = \sum k{:} nodes\ k\ splits\ on\ feature\ j^{S_kC_k} - (5)$$

"fjj" denotes the significance of feature "j."

"s$_{\text{k}}$" represents the quantity of samples that reach node "k."

"Ck" indicates the impurity value linked to node "k."

To calculate the collective feature importance on the Random Forest level, the first step involves normalizing the feature importance for each tree with respect to that specific tree:

$$normsfj_j = \frac{fj_j}{\sum k \in \mathrm{all}\ fetures^{\ fj_k}} ----(6)$$

"normfij" signifies the normalized importance of feature "j".

"fj sub(j)" signifies the importance of feature "j".

Next, The importance values of features from each tree are added together and subsequently normalized.

$$RF\ fj_j = \frac{\sum k\ normsfj_{jk}}{\sum k \in \text{all features}, l \in \text{all } trees^{\,normsfj_{kl}}} - -(7)$$

"RFfij" represents the feature importance of feature "j" computed across all the trees within the Random Forest model.

"normfjjk" corresponds to the normalized feature importance of feature "j" within the context of tree "k" in the Random Forest.

ML techniques works with machine to learning and trained with datasets or specific pattern to quickly and easily find any disease risk in any patients with feature extraction and some classification models and algorithms. In this work, we use ML models to predict and classify diabetes disease patients, here datasets are extracted with kaggle datasets and using some SML models like KNN, DT, SVC, LR, GNB, and RF, All models were applied to a pre-processed dataset, and performance parameters such as accuracy, precision, recall, and F1-score were determined using a test dataset.

This works main aim is to find and detect diabetes patients at an early stage with accurate and quickly mode. As we know that any disease is affected not only to patients but also their family and each family persons, and diabetes is a diseases some time it may lead to patient’s death also. Our model predicts and compares the different study and presents the results summary with performance parameters and also shows better results than existing techniques or models.

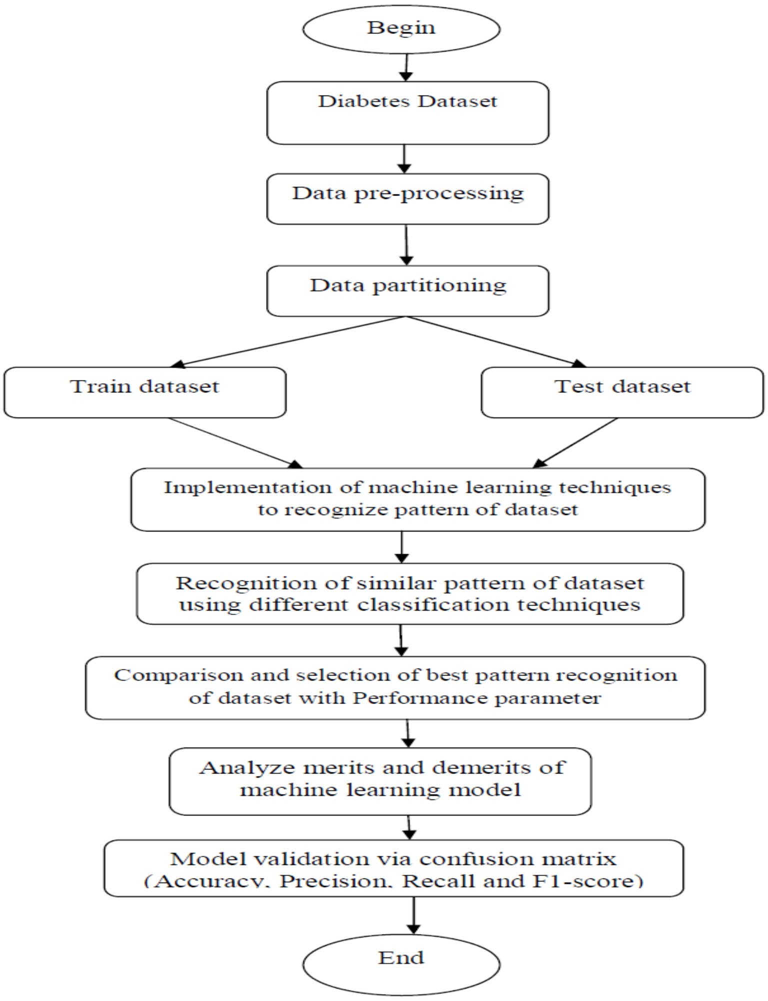
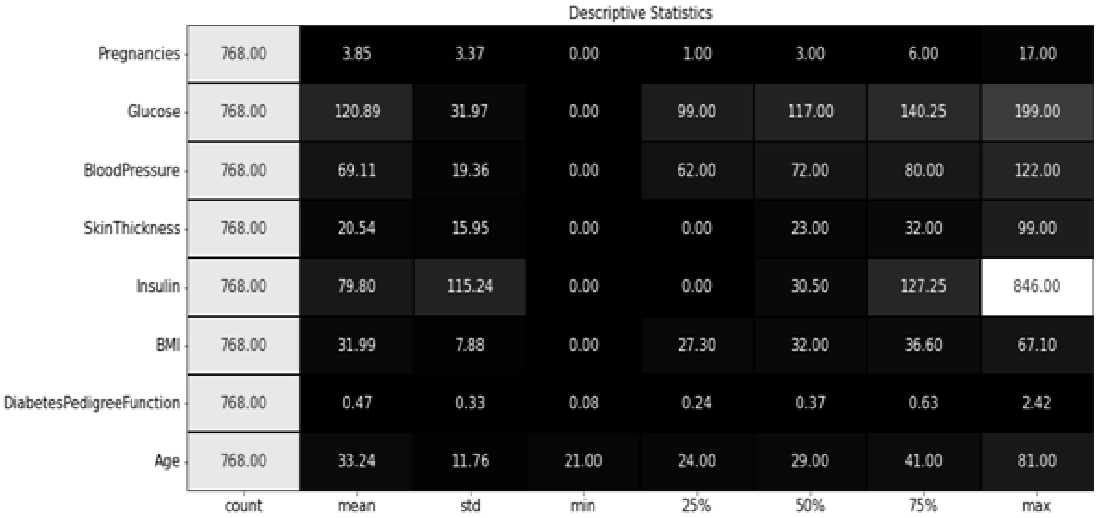

Begin
Diabetes Dataset
Data pre-processing
Data partitioning
Train dataset
Test dataset
Implementation of machine learning techniques to recognize pattern of dataset
Recognition of similar pattern of dataset using different classification techniques
Comparison and selection of best pattern recognition of dataset with Performance parameter
Analyze merits and demerits of machine learning model
Model validation via confusion matrix (Accuracy, Precision, Recall and F1-score)
End

Figure 3: Diabetes disease prediction system with ML techniques flow graph [8].



Descriptive Statistics

	count	mean	std	min	25%	50%	75%	max
Pregnancies	768.00	3.85	3.37	0.00	1.00	3.00	6.00	17.00
Glucose	768.00	120.89	31.97	0.00	99.00	117.00	140.25	199.00
BloodPressure	768.00	69.11	19.36	0.00	62.00	72.00	80.00	122.00
SkinThickness	768.00	20.54	15.95	0.00	0.00	23.00	32.00	99.00
Insulin	768.00	79.80	115.24	0.00	0.00	30.50	127.25	846.00
BMI	768.00	31.99	7.88	0.00	27.30	32.00	36.60	67.10
DiabetesPedigreeFunction	768.00	0.47	0.33	0.08	0.24	0.37	0.63	2.42
Age	768.00	33.24	11.76	21.00	24.00	29.00	41.00	81.00

Figure 2: This picture represent dataset statics for experimental work.

IV. EXPERIMENTAL RESULT

To assess the success of the planned research, this study employs a range of performance evaluation criteria [1]. Diabetes can significantly impact life expectancy and quality of life, making early prediction of this chronic disorder crucial for reducing long-term risks and complications [2]. The performance metrics used for evaluation are as follows: Accuracy (Ac), Precision (Pe), Recall (Re), and F1 Score (F1-s) [4].

Accuracy: Usually a measure of how often classifiers are estimated correctly. The accuracy rate of the developed algorithm is calculated according to Equation.

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

Sensitivity (Recall): Indicates the rightly classified positive values from the positive samples' total count.

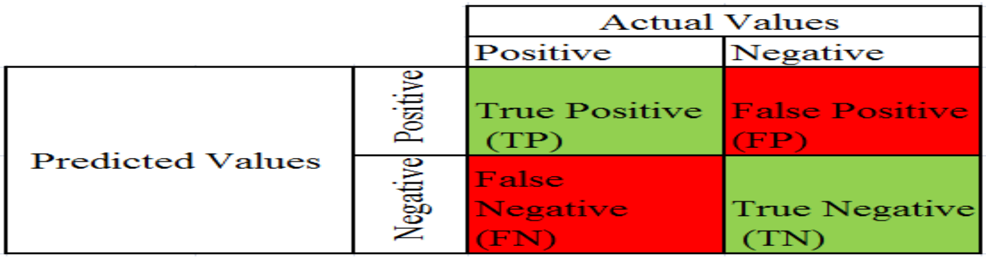
$$\text{Recall}=\frac{\text{TP}}{\text{TP+FN}}$$

Precision: It shows predicted and actual predicted value ratio.

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

F1 Score: Precision and sensitivity are the harmonic means, so it's a powerful metric for calculating model performance. The precision and recall in calculating the F1-Score is represented respectively.

$$\text{F1Score}=\frac{2*\text{Precision}*\text{Recall}}{\text{Precision+Recall}}$$



		Actual Values	
		Positive	Negative
Predicted Values	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Figure 4: Performance evaluation using confusion matrix [11].

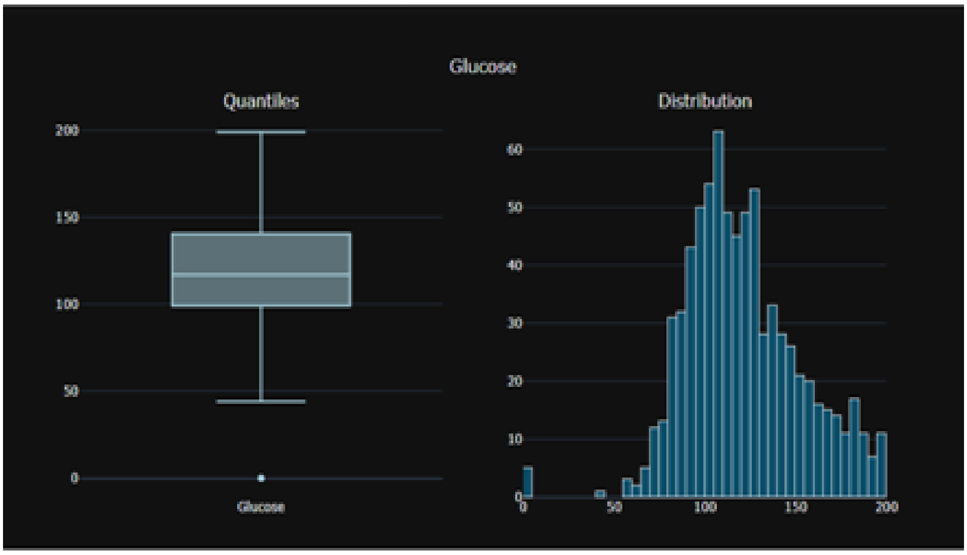


Figure 5: This picture represents glucose features quantities and distribution study for experimental work.

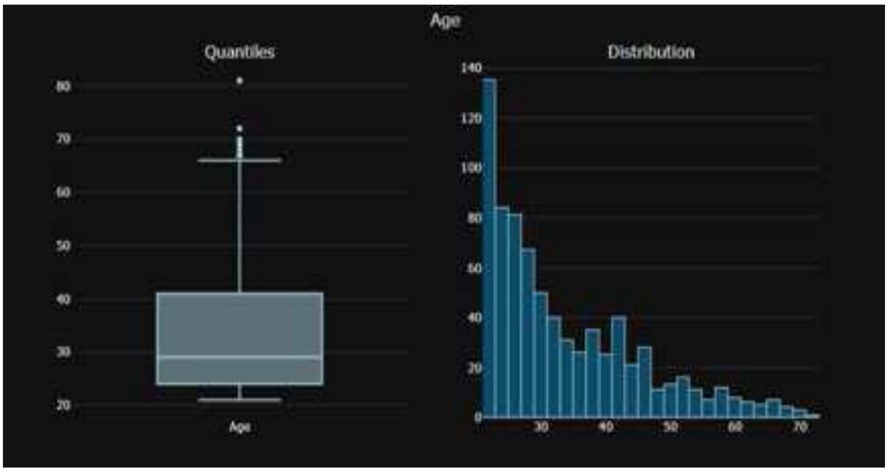


Figure 6: This picture represents age features quantities and distribution study for experimental work.

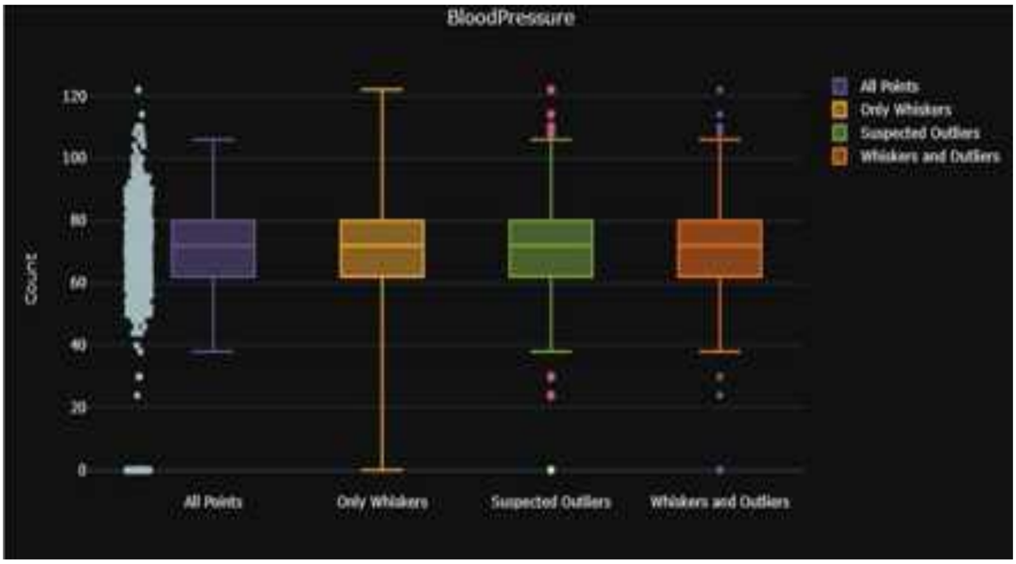


Figure 7: This picture represents Blood Pressure features quantities and distribution study for experimental work

After applying all the ML models finally we found some performance parameters value ad all models are compared with their respective performance parameters value, here we present some model experimental summary and their comparative table also.

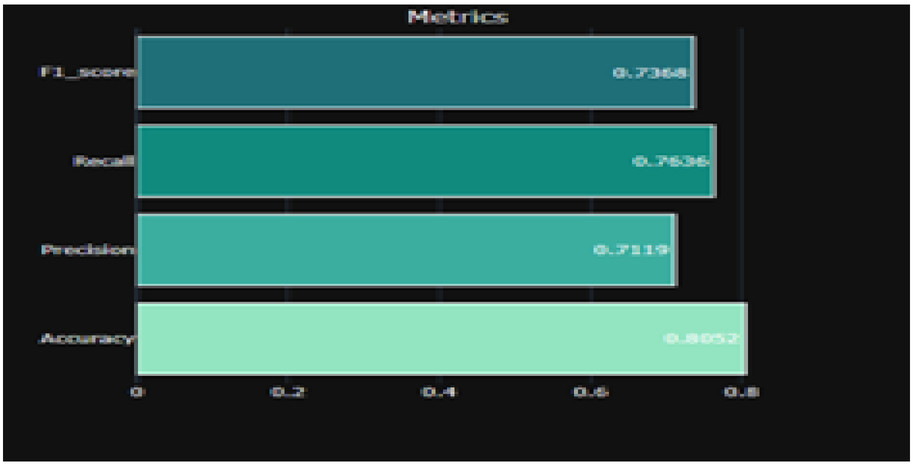


Figure 8: This picture represents the performance parameter matrix of the random forest model for experimental work.

Machine Learning Model	F1-Score	Recall	Precision	Accuracy
Our model	0.83	0.85	082	0.87
KNN	0.76	0.78	0.74	0.83
DT	0.73	0.76	0.71	0.80
GNB	0.69	0.92	0.56	0.70
LR	0.75	0.81	0.70	0.81
SVC	0.76	0.82	0.73	0.82

Table 1: The above table shows comparative study of different machine learning model.

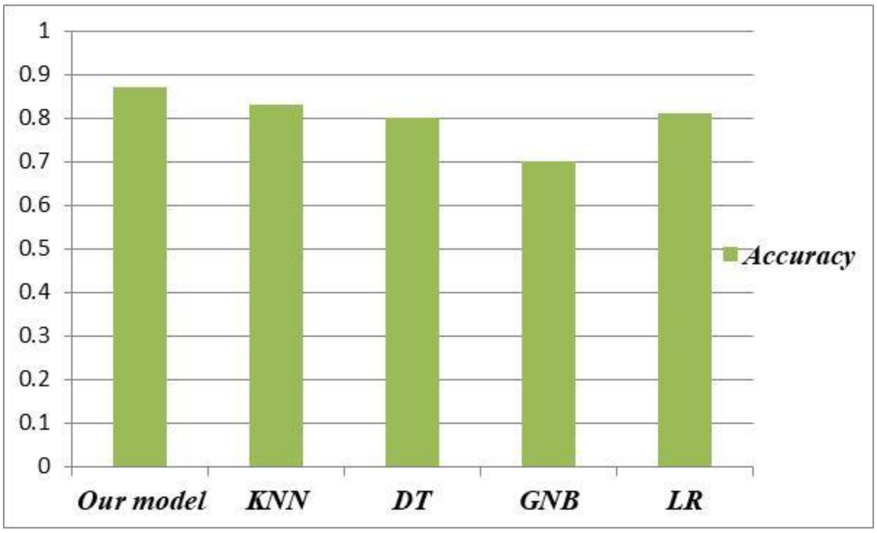


Figure 9: This picture represents the performance parameter value of accuracy between different machine learning techniques.

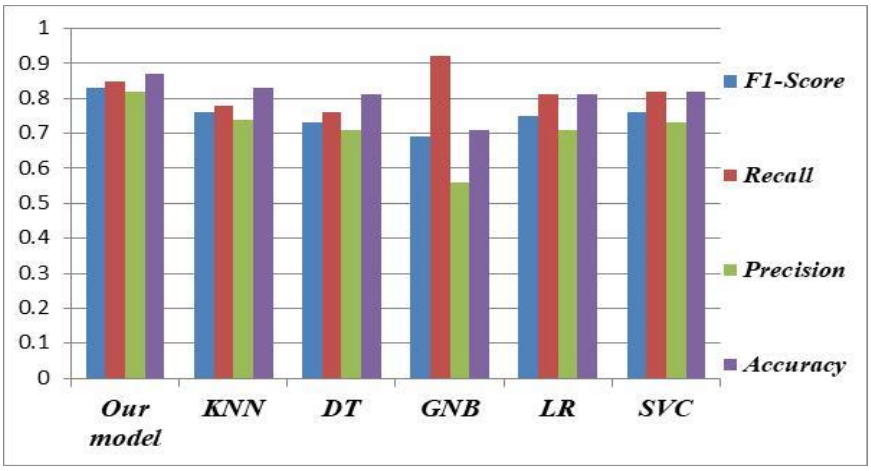


Figure 10: This picture represents the performance parameter value of accuracy, F1-Score, precision, and recall between different machine learning techniques.

V. Conclusion

Machine learning algorithms are gaining popularity in various fields of bioinformatics, medical imaging, and biomedical research, providing promising performance and aiding medical professionals in early diagnosis. In this study, the open-source Pima Indian dataset has been utilized. To address the issue of imbalanced class problems, SMOTE preprocessing techniques have been applied. Future research can explore the integration of additional algorithms, such as deep neural networks, to further enhance accuracy and precision. Additionally, the use of swarm optimization techniques can be considered to optimize results. Application program development could also be incorporated to enhance the overall system.

References

[1] Muhammad ExellFebrian, FransiskusXaveriusFerdinan, “Diabetes prediction using supervised machine learning”, 7th International Conference on Computer Science and Computational Intelligence ,2023, pp. 21-30.

[2] Kok-Lim Alvin Yau, Yung-Wey Chong, “Reinforcement Learning Models and Algorithms for Diabetes Management”, IEEE Access, 2023, pp. 28391-28415.

[3] NavaneethBhaskar, VinayakBairagi, “Automated Detection of Diabetes From Exhaled Human Breath Using Deep Hybrid Architecture”, IEEE Access, 2023, pp. 51712-51723.

[4] Abdul MuizFayyaz, Muhammad Imran Sharif, “Analysis of Diabetic Retinopathy (DR) Based on the Deep Learning”, Information 2023, pp. 1-14.

[5] MehnoorAhsan, SaeedaNaz, “A Deep Learning Approach for Diabetic Foot Ulcer Classification and Recognition”, Information 2023, pp. 1-11.

[6] D Rathore, PMannepalli, “Diseases Prediction and Classification Using Machine Learning Techniques”, AIP Conference Proceedings 2424, 070001 (2022); https://doi.org/10.1063/5.0076768.

[7] Maria Tariq, Vasile Palade, “Diabetic Retinopathy Detection using Transfer and Reinforcement Learning with effective imagepreprocessing and data augmentation techniques”, 2023, pp. 1-30.

[8] IsfafuzzamanTasin, TansinUllah Nabil, “Diabetes prediction using machine learning and explainable AI Techniques”, Healthcare Technology Letters, wiley, 2023, pp. 1-10.

[9] Reema Shah, Jeremy Petch, “Nailfoldcapillaroscopy and deep learning in diabetes”, wiley, 2023, pp. 145-151.

[10] ErdalOzbay, “An active deep learning method for diabetic retinopathy detection in segmented fundus images using artificial bee colony algorithm’, Artificial Intelligence Review, 2023, pp. 3291-3318.

[11] Mahesh S Patil, “Effective Deep Learning Data Augmentation Techniques for Diabetic Retinopathy Classification”, International Conference on Machine Learning and Data Engineering, 2023, pp. 1156-1165.

[12] PoshamUppamma, Sweta Bhattacharya, “Diabetic Retinopathy Detection: A Blockchain and African Vulture Optimization Algorithm-Based Deep Learning Framework”, Electronics 2023, pp. 1-19.

[13] D Rathore, PKMannepalli, “A Review of Machine Learning Techniques and Applications for Health Care “, International Conference on Advances in Technology, Management & Education, 2021, IEEE proceeding, 978-1-7281-8586-6/21.

[14] Huiqi Lu,Xiaorong Ding, “Digital Health and Machine Learning Technologies for Blood Glucose Monitoring and Management of Gestational Diabetes”, IEEE, 2022, pp. 1-19.