

PERFORMANCE EVALUATION OF DATA HANDLING ALGORITHMS (DT, KNN, RF AND DNN) UNDER THE HANDLING OF TWO DATASET OF DIABETES

Dr. Vinay Bhatt*¹ Prof. (Dr.) Ajay Kumar Mishra², Dr. Vivek Bhatt³ Prof (Dr.) Praveen Kumar⁴,

¹Assistant Professor, Department of Computer Science, Shri Khushal Das University (SKDU), Hanumangarh, Rajasthan

² Professor, Department of Mathematics, Sikkim Skill University, Namthang, Sikkim

³Associate Professor, Department of Shalya Tantra, Smt. Manjira Devi Ayurvedic Medical College & Hospital, Uttarkashi, Uttarakhand

⁴Professor, Department of Computer Science and Engineering, Asian International University, Imphal West, Manipur

***Corresponding author email:**

vinay10191@gmail.com

Abstract

Diabetes is a condition that affects people of all ages these days, as everyone is aware. This paper discusses the implementation of algorithms based on two separate diabetes datasets. The algorithms were chosen based on the highest value and accuracy. In this paper analyze the four algorithms such as DT, KNN, RF and DNN. The main objective of this paper is to handle data with different and large datasets because data is the most difficult to handle. In the previous research, diabetes was predicted in different cases and their conclusions were also different. In this paper, the DNN algorithm is used and compared with other algorithms under two different datasets. In this paper, compare the DNN algorithm with the proposed algorithms such as DT, RF and KNN on large and two different datasets of diabetes. The first dataset is taken from Pima Indians Diabetes Database (PIDD) and the second dataset is collected through survey from Manjira Devi Hospital Uttarkashi (MDHU). On the basis of these two datasets, algorithms are compared.

Keywords: Diabetes Prediction, PIDD, MDHU, KNN, DT, RF, DNN

I. Introduction

The term "diabetes" describes a collection of metabolic diseases marked by high blood glucose levels brought on by either inadequate insulin synthesis, poor insulin use, or a combination of the two [12][14]. Long-term harm and malfunction of organs like the heart, blood vessels, kidneys, eyes, and nerves are associated with chronic hyperglycemia [16][18]. Depending on their chronological age, income, color, and ethnicity, people with diabetes experience different affects [10]. Diabetes can be caused by both hereditary and environmental causes, which lead to resistance to insulin and beta-cell death [15][17]. In this work, evaluate the DNN method on two distinct and sizable datasets to earlier top techniques including DT, RF, and KNN. The Pima Indians Diabetes Database (PIDD) provided the first dataset, while Manjira Devi Hospital Uttarkashi (MDHU) provided the second dataset via survey.

II. Literature Survey

In this section present the review of literature on behalf of previous research showing in table 1 as:

Author's	Years	Research Contributions
El-Sayyid El-Bashbishy, A., & El-Bakry, H. M.	2026	Proposed the study on evaluation of pediatric diabetic prediction using nine machine learning algorithms. The outcome of this study is ANN is superior to other proposed algorithms [1].
Ayoade, O. B., Shahrestani, S., & Ruan, C.	2025	Proposed the performance comparison between ML and DL for diabetes prediction progression. The outcome of this evaluation is deep learning model is outperforming than machine learning models [2].
Zaferani, N., Afrash, M. R., & Moulaei, K.	2025	Proposed the transparent ensemble model by combining the three algorithms as KNN, NN and random forest for the prediction of type-2 diabetes. The result of this research is random forest gain the 100 % accuracy than other algorithms [3].
Hameed, E. M., Joshi, H., & Kadhim, Q. K.	2025	Proposed the review on advancement of AI technique for diabetes prediction. The outcome of this review is DNN and decision tree algorithm gains the best accuracy to other algorithms [4].
Cveticanin, L., & Arsenovic, M.	2025	Proposed the review on predictive model of diabetes for children by using machine learning and deep learning model. The outcome of this review is deep learning model is easily diagnosed to diabetes for children [5].
Altamimi, A.et al.	2024	Proposed the automated approaches of machine learning for predict to diabetes by using KNN algorithm of ML with data mining techniques. The result of this work is KNN is higher accurate algorithm [7].

Table 1. Review of Literature

III. Design of System for Handle the Data

Therefore, the objective of this paper is to handle large dataset of diabetes for which algorithms like DT, KNN, RF and DNN are used. Let's compare all these algorithms under two different diabetes datasets. This work is done through SMOTE technique in which classes are selected and algorithms are used properly so that evaluation and conclusions are generated well [11][13]. Figure 1 below shows the design of the research in which the functionality of the algorithms is explained and how the data is analyzed after collecting the dataset. First of all, two different datasets are collected and preprocessed, after which attributes are extracted from them. In the next step, the data is divided, after which the data is analyzed through the sampling technique SMOTE and finally after the analysis, the findings are assessed.

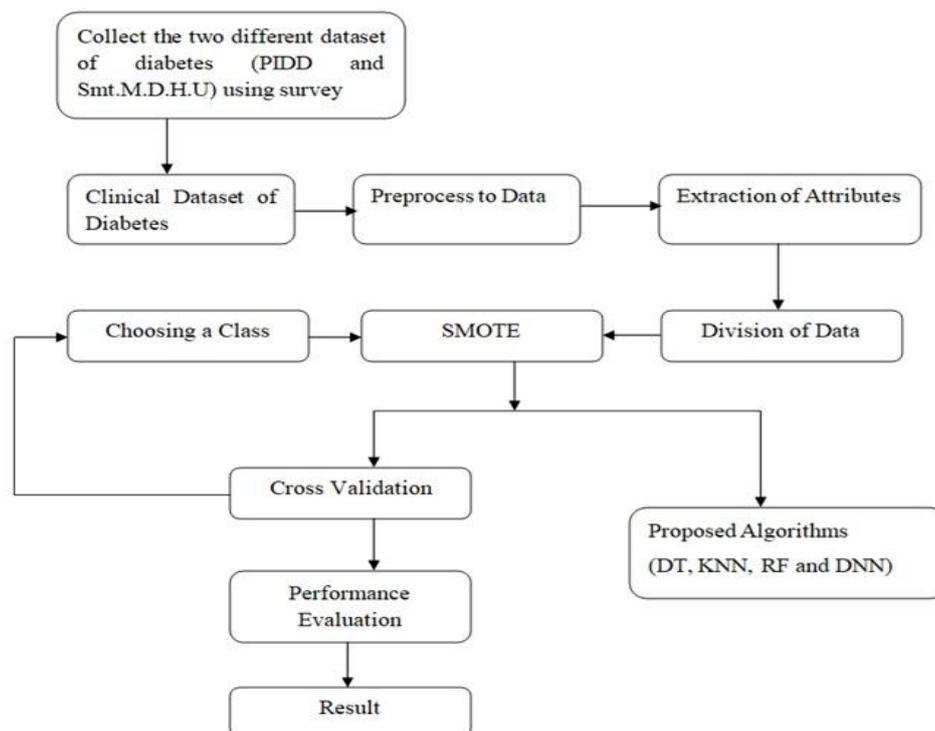


Figure 1. System design under data handling by proposed algorithms

IV. Scaling of Data

PIDD and M.D.H.U are the sources of the primary and secondary datasets, respectively, that are collected in this paper. When both datasets have been collected, they are correctly scaled based on the attributes. To ensure effective testing, the dataset is then trained for analysis. Both scale the dataset according to 9 attributes and prepare it for implementation. Table 2 shows the diabetes dataset along with its attributes.

Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
7	149	71	34	0	32.3	0.626	51	1
1	86	67	29	0	26.7	0.352	32	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0
10	168	74	0	0	38	0.537	34	1
10	139	80	0	0	27.1	1.441	57	0
1	189	60	23	846	30.1	0.398	59	1
5	166	72	19	175	25.8	0.587	51	1

Table 2. Dataset of diabetes

V. Proposed Algorithm

This paper handles diabetic data using four algorithms: DT, KNN, RF, and DNN.

- i. *DT*: A collection of attributes, including pregnancies, glucose levels, skin thickness, insulin, body mass index and age are used to create a decision tree [6][8]. A dataset of individuals who have or not diabetes is used to train the tree. The probability of diabetes in newly diagnosed individuals is then calculated using the tree.
- ii. *KNN*: Using the dataset's closest k-neighbors, this non-parametric algorithm creates predictions [3][8]. When diagnosing diabetes, it may be applied to correct data that is missing. The result may increase the predicted diabetes accuracy.
- iii. *RF*: The findings of multiple algorithms are combined in such a form of combined learning technique [3][9]. According to the research, the algorithm for the Random Forest approach can enhance the effectiveness of diabetes initial detection.

iv. *DNN*: One kind of ANN that is helpful for accurately analyzing vast volumes of data is the DNN [2][10]. An algorithm successfully predicts diabetes using numerous times and ten layers that are hidden.

VI. Performance Evaluation and Results

Plots and graphs in figure 2, 3, and 4 can be created using python tools that are applied based on attributes, and the four algorithms mentioned above are used to evaluate data.



Figure 2.pyplot of simulation

Analyze the results using a Python tool after examining the data in figures 2, 3, and 4. The Pyplot is predicted in figure 2, the algorithm's prediction is shown in the correlation matrix by the attributes in figure 3, and the column chart is generated in figure 4.

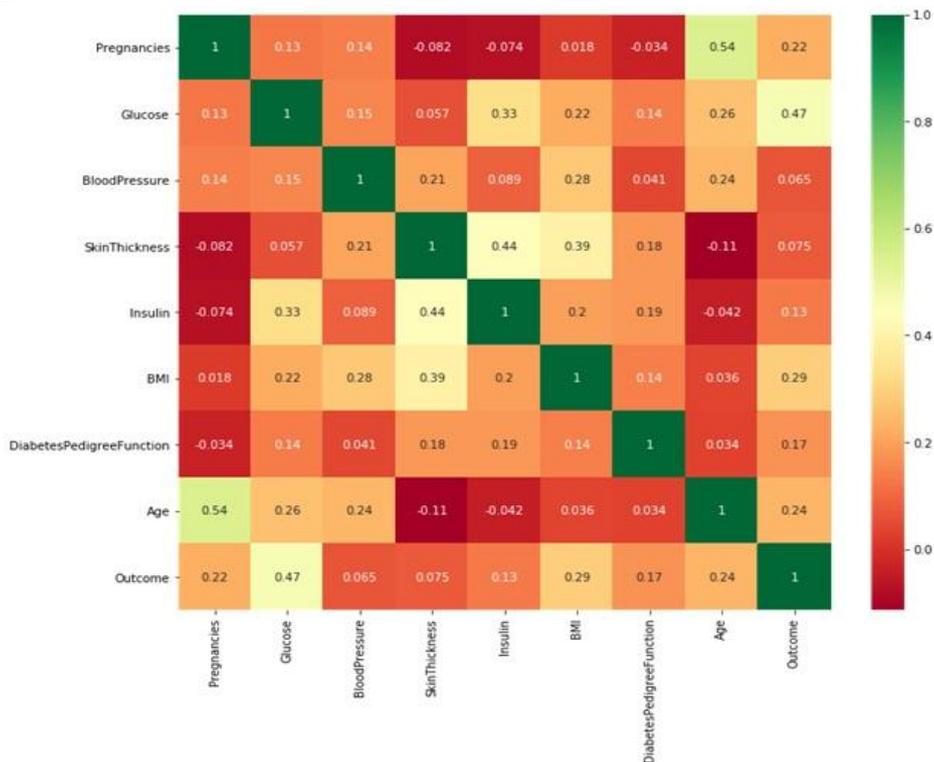


Figure 3. Correlation Matrix

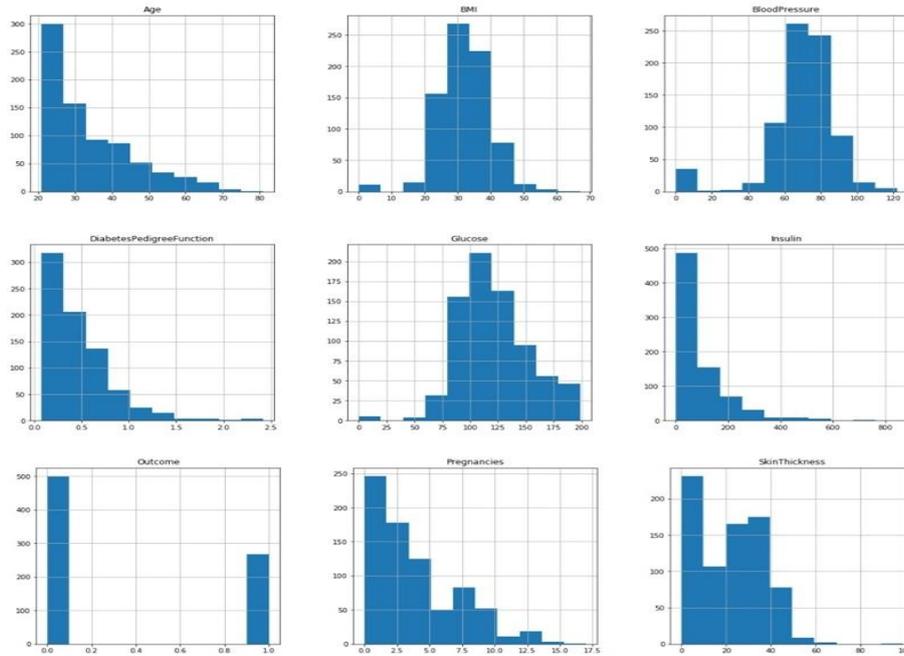


Figure 4. Chart of prediction

VII. Performance Parameters

In this chapter, diabetes is predicted using algorithms based on the four metrics listed below.

In diabetes observation, if there are four values—true positive (a1), true negative (a2), false positive (a3), and false negative (a4), then

i. Accuracy (A_c) : $A_c = \frac{a1+a2}{a1+a2+a3+a4}$

ii. Precision (P_c): $P_c = \frac{a1}{a1+a3}$

iii. Recall (R_c): $R_c = \frac{a1}{a1+a4}$

iv. F1 Score ($F_1\ score$): $F_1\ score = 2 * \frac{P_c * R_c}{P_c + R_c}$

VIII. Outcomes of Evaluation

The results are assessed using the dataset 1 (PIDD) and the dataset 2 (MDHU) in table 3 and table 4. Following the analysis of both datasets, all four algorithms (DT, KN, RF, and DNN) consider the findings of both. The four algorithm performance metrics mentioned above serve as the foundation for this conclusion. The second table 4 is predicted by algorithms based on the MDHU dataset, while the first table 3 below is based on the PIDD Dataset.

Algorithms	A_c	P_c	R_c	$F_1\ score$
DT	0.9034	0.7062	0.6531	0.7213
KNN	0.8901	0.4190	0.8210	0.6534
RF	0.9113	0.4621	0.7120	0.5123
DNN	0.9536	0.8101	0.8438	0.8654

Table 3. PIDD dataset

Algorithms	A_c	P_c	R_c	$F_1\ score$
DT	0.8207	0.6201	0.4301	0.7013
KNN	0.7404	0.3103	0.7201	0.5801
RF	0.8380	0.3045	0.6802	0.4702
DNN	0.9030	0.7801	0.7103	0.7245

Table 4. MDHU dataset

IX. Discussion of Comparison

The results are described in detail in this section. The algorithms (DT, KNN, RF, and DNN) are compared in the figures below based on how well they perform on the two datasets, MDHU and PIDD, respectively. The results are displayed in the tables above. When all of the methods are compared on the accuracy parameter between two datasets, Figure 5 shows that the DNN algorithm has the best accuracy and the KNN algorithm has the lowest.

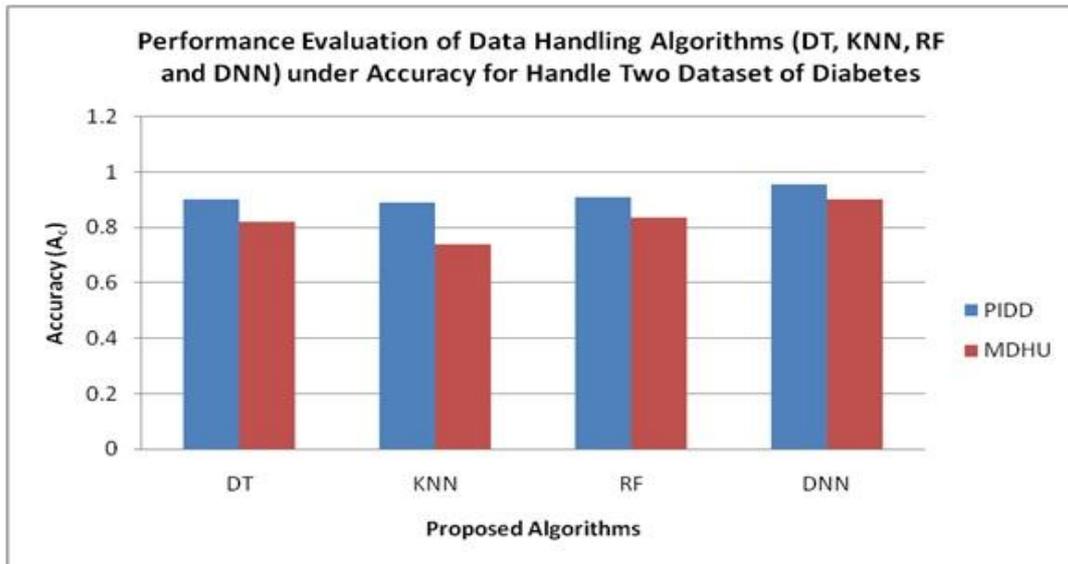


Figure 5. Accuracy prediction by algorithms

The algorithms will be compared using the precision parameter in figure 6, where DNN has the best score and KNN has the lowest score.

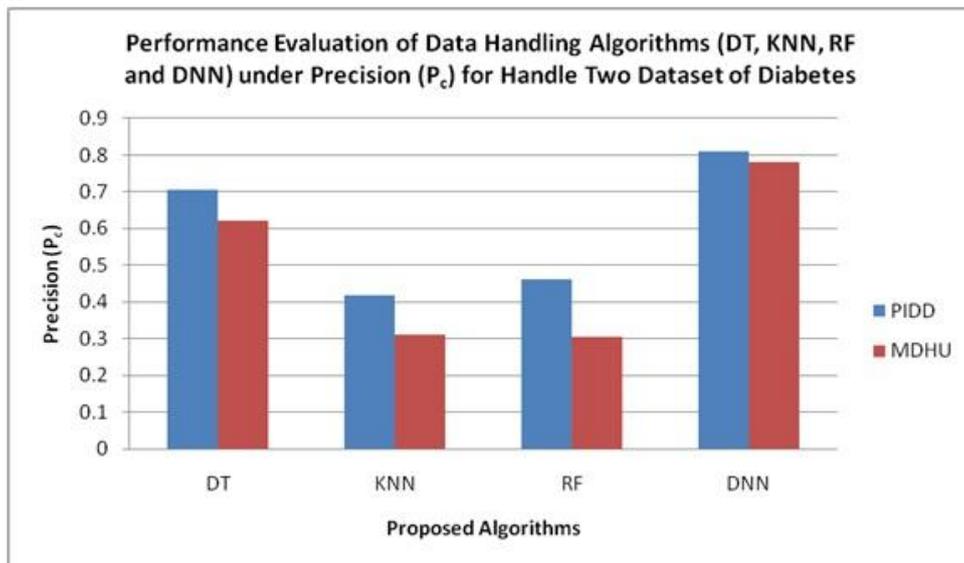


Figure 6. Comparison under precision

Figure 7 compares the algorithms based on recall parameters, with DNN having the highest recall value and DT having the lowest.

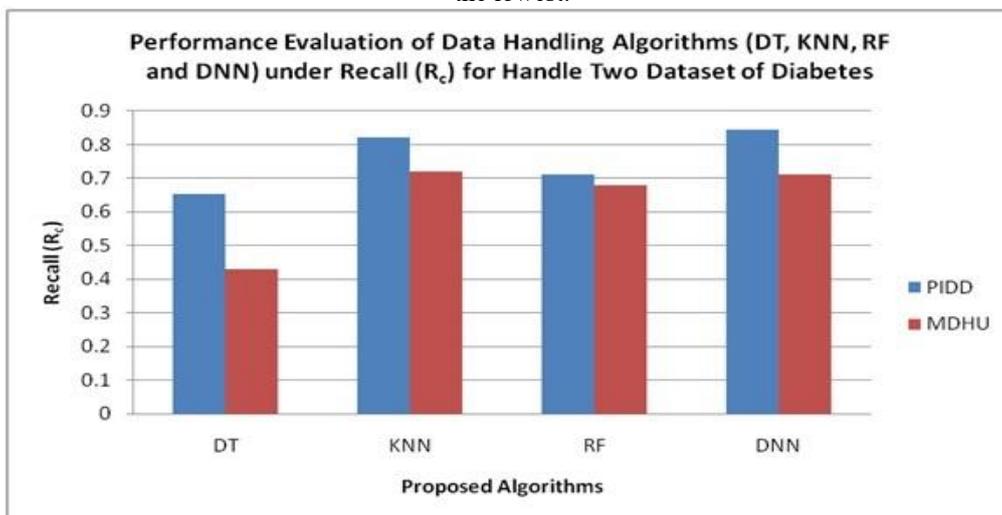


Figure 7. Recall prediction by algorithms

In the fourth figure 8, the algorithms are compared using the F1 Score, with RF representing the lowest and DNN being the greatest.

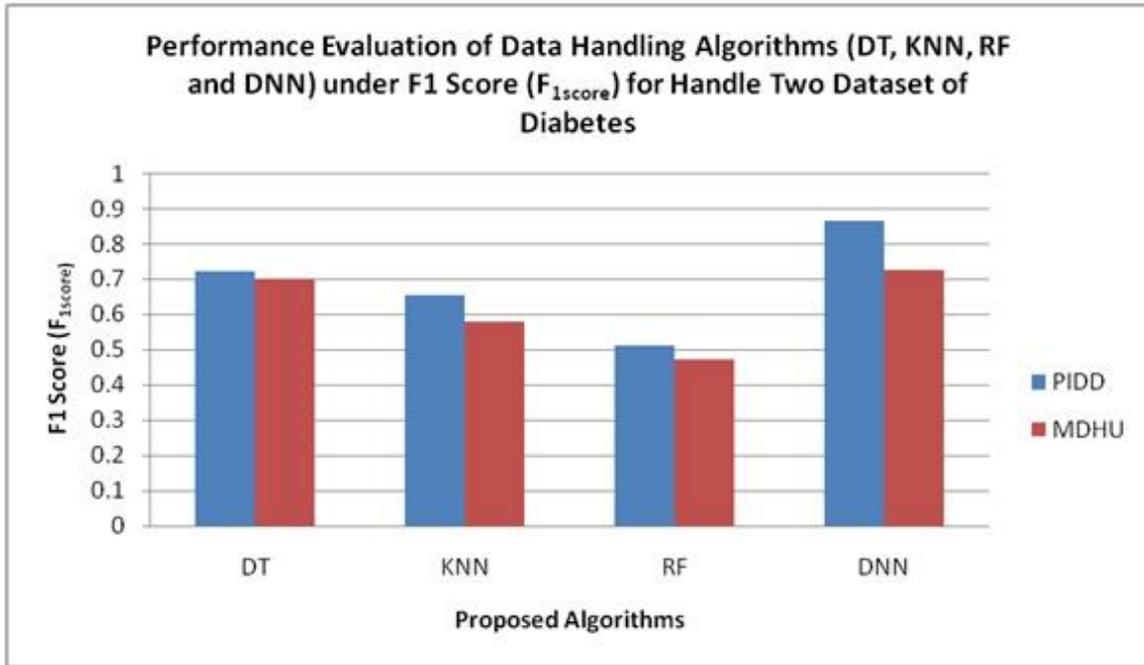


Figure 8. Evaluation under F1 Score

X. Conclusion

Diabetes is extremely difficult to track and handle due to its enormous dataset, which is the primary goal of this paper. In this paper, four machine learning algorithms—DT, RF, KNN, and DNN—have been used to predict diabetes. Two distinct datasets, such as Pima Indian Diabetes Database (PIDD) and Manjira Devi Hospital Uttarkashi (MDHU), have been used in research based on nine attributes to predict diabetes. Four performance metrics—accuracy, precision, recall, and F1 score—are used to compare the algorithms' performances. The complete analysis concludes that, when compared to other algorithms like RF, KNN, and DT, the DNN algorithm is the most successful at handling diabetes datasets. Further work of this paper is implementation of DNN algorithm in more cases of diabetes and other diseases in healthcare systems.

References

1. El-Sayyid El-Bashbishy, A., & El-Bakry, H. M. (2026). Pediatric diabetes prediction using machine learning. *Scientific Reports*, 16(1), 1979.
2. Ayoade, O. B., Shahrestani, S., & Ruan, C. (2025). Machine learning and deep learning approaches for predicting diabetes progression: A comparative analysis. *Electronics*, 14(13), 2583.
3. Zaferani, N., Afrash, M. R., & Moulaei, K. (2025). Predicting and classifying type 2 diabetes using a transparent ensemble model combining random forest, k-nearest neighbor, and neural networks. *Scientific Reports*.
4. Hameed, E. M., Joshi, H., & Kadhim, Q. K. (2025). Advancements in artificial intelligence techniques for diabetes prediction: a comprehensive literature review. *Journal of Robotics and Control (JRC)*, 6(1), 345-365.
5. Cveticanin, L., & Arsenovic, M. (2025). Prediction models for diabetes in children and adolescents: A review. *Applied Sciences*, 15(6), 2906.
6. Rehman, T. U., Iqbal, J., Roman, M., Ahmad, Z., Ahmad, B., & Rafiq, M. (2025). A comparative machine learning approach to diabetes prediction: integrating best first search feature selection with svm and naïve bayes classifiers. *Spectrum of Engineering Sciences*, 3(12), 37-49.
7. Altamimi, A., Alarfaj, A. A., Umer, M., Alabdulqader, E. A., Alsubai, S., Kim, T. H., & Ashraf, I. (2024). An automated approach to predict diabetic patients using KNN imputation and effective data mining techniques. *BMC Medical Research Methodology*, 24(1), 221.
8. Musa, J., & Abdulazeez, A. M. (2024). A review on diabetes classification based on machine learning algorithms. *The Indonesian Journal of Computer Science*, 13(2).
9. Mahmud, S., Islam, B. U., Anik, N. H., & Ghosh, T. (2024, September). Diabetes Prediction: A Comparative Analysis of Machine Learning Algorithms with SMOTE. In *2024 IEEE International Conference on Computing, Applications and Systems (COMPAS)* (pp. 1-6). IEEE.
10. Alnowaiser, K. (2024). Improving healthcare prediction of diabetic patients using KNN imputed features and tri-ensemble model. *IEEE Access*, 12, 16783-16793.
11. Alketbi, S., Albadawi, A., & Al Aghbari, Z. (2024, June). A deep learning-based ensemble method for early detection of diabetes. In *2024 Advances in Science and Engineering Technology International Conferences (ASET)* (pp. 01-06). IEEE.

12. Ramya, S., & Kalaivani, D. (2023, December). Analysis of Machine Learning and Deep Learning for Diabetes Diagnosis in Design Thinking. In *2023 International Conference on Energy, Materials and Communication Engineering (ICEMCE)* (pp. 1-7). IEEE.
13. Negi, P., Dhablia, A., Vanjari, H. B., Tamkhade, J., Ikhar, S., & Shirkanade, S. T. (2023, November). Evaluating feature selection methods to enhance diabetes prediction with random forest. In *Proceedings of the 5th International Conference on Information Management & Machine Intelligence* (pp. 1-7).
14. Bhat, S. S., Selvam, V., & Ansari, G. A. (2023). Predicting life style of early diabetes mellitus using machine learning. *International Journal of Computing*, 22(3-4), 345-351.
15. Salem, H., Shams, M. Y., Elzeki, O. M., Abd Elfattah, M., F. Al-Amri, J., & Elnazer, S. (2022). Fine-tuning fuzzy KNN classifier based on uncertainty membership for the medical diagnosis of diabetes. *Applied Sciences*, 12(3), 950.
16. Afsaneh, E., Sharifdini, A., Ghazzaghi, H., & Ghobadi, M. Z. (2022). Recent applications of machine learning and deep learning models in the prediction, diagnosis, and management of diabetes: a comprehensive review. *Diabetology & Metabolic Syndrome*, 14(1), 196.
17. Olisah, C. C., Smith, L., & Smith, M. (2022). Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective. *Computer Methods and Programs in Biomedicine*, 220, 106773.
18. Sabitha, E., & Durgadevi, M. (2022). Improving the diabetes diagnosis prediction rate using data preprocessing, data augmentation and recursive feature elimination method. *International Journal of Advanced Computer Science and Applications*, 13(9).