International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Evaluate the Predictive Performance of Supervised Machine Learning Algorithms in Diabetes Dataset

Md. Jamaner Rahaman

Leading University, Department of Computer Science and Engineering, Sylhet, Bangladesh
Email: *jamaner.rahaman[at]gmail.com*

Abstract: ***The progression of diabetes disorder is very alarming in our society nowadays. Not only increase the glucose level of a human body but also spread other complexities like blood pressure, heart disease, kidney disease and many more diseases that have the relation with diabetes. Still now most of the people are not aware of doing diagnosis and sometimes it is hard to find early also. Diagnosis of diabetes is more important to prevent or control this disease, so that machine learning might play a very significant role in this area because of its less error and efficiency. In order to predict diabetes, the author of this research attempted to evaluate the effectiveness of several machine learning (ML) methods, including Random Forest, SVM, Decision Tree, K-Nearest Neighbors, and Logistic Regression with the help of Pima Indians Diabetes Dataset from Kaggle repository. For each model confusion matrix has been used to measure Accuracy, Precision, Recall, and F1-Score for evaluating the performance. Random Forest got the highest accuracy of 81% while Decision Tree got the least accuracy of 69%. The methodology proposed in this paper is very easy to understand for everyone which is one of the major key points of this research.***

Keywords: Diabetes, Random Forest, Machine Learning, Supervised Learning

1. Introduction

There are many diseases in the human body but some common diseases nowadays we can see, diabetes is one of them. Increasing the glucose or sugar level in the blood flow of a human body responsible for diabetes. A hormone called insulin found from pancreas is maintaining the glucose level but when the production of insulin is not enough compared to the glucose level then glucose molecules are not able to enter the cell so that energy being reduced and glucose remain in the blood flow [1]. Kidney failure, blindness, high blood pressure, heart failures these are the top most side effects of diabetes. The American Diabetes Association said that on the basis of a statistical report 3 Crore people were suffering with diabetes in 2015. 42 Crore people suffering from diabetes are being told by the World Health Organization all over the world. Type-1, Type-2, and gestational diabetes are the three varieties of the disease. Type-1 diabetes occurs because of low secretion of insulin so that the patient needs to take insulin daily. Fatigue, constant hunger, unexplained weight loss, thirst and constant urination are the symptoms of Type-1 diabetes. Type-2 diabetes occurs because of how much insulin is produced by the pancreas which the body is unable to use effectively. Increased body weight and less exercise are the main causes of Type-2 diabetes also it is hard to find. Gestational diabetes occurs during the woman's pregnancy and once the delivery is over then it is recovered [2], [3], [4].

Diabetes is not so harmful when it is controlled. The best way to control diabetes is early detection or diagnosis. In that sense, with the help of technology we can easily predict diabetes early. Nowadays machine learning plays a vital role in the healthcare system to identify various diseases efficiently with less error. At first the machine learning algorithm learns from the data then acts which it has learnt. Dataset is one of the major components for predicting correctly because larger the dataset and well organized dataset will produce the better result. In the healthcare system large amounts of data are already stored because of their huge number of patients [5]. So, statistical analysis using machine learning is quite easy and efficient in this field at the same time proper algorithms need to be selected. Which dataset is suited for which algorithm this is also the criteria to get better performance. First take the dataset then pre-processed data implemented with the help of algorithms and finally evaluate the algorithms based on their result. There are several types of machine learning algorithms we can see like supervised, un-supervised, semi-supervised and reinforcement. In this paper the author tried to analyze the performance of five commonly used supervised machine learning algorithms on diabetes dataset and select the best one. People can easily get the idea of the best performing algorithm from this research and use it in their work especially in the healthcare system to reduce the diabetes disorder.

The following few sections like: Section 2 will describe literature review, sections 3 and 4 will describe methodology and result analysis including discussion, finally section 5 will describe the conclusion part.

2. Literature Review

There are many researchers already doing so much study in the area of the healthcare system with the help of machine learning among them some of the related works are given below.

Kakoly et al. [6] basically worked with two-fold feature selection techniques such as PCA, IG for increasing the accuracy level. Processed features applied into five most popular commonly used ML algorithms (DT, SVM, RF, LoR, KNN). They collected 738 data from Bangladesh by using survey questions of which 256 participants had diabetes and 482 nondiabetic participants. They divided the

dataset by 80:20 ratio and finally got an accuracy of 82.2%.

In [7], the authors mainly used feature selection methods to select the suitable feature rather than all features at the same time they performed data cleaning process. To find out the best result they also worked with some existing approaches. Logistic regression showed the best accuracy of 85 % with selected features compared to other models.

Hirnak et al. [8] proposed a method based on extracting the attributes which gives the result in early detection and also proposed deep learning approaches. Existing systems which are available right now, they did an analysis on them. They used three datasets while the third one was the combination of the first two datasets. In their research, a Random Forest algorithm gave the best result and they also developed a website.

Vakil et al. [9] did a comparative analysis on several machine learning algorithms focusing on feature attribution to identify the most important feature by using SHAP. They collected data from Sylhet Diabetes Hospital in Sylhet, Bangladesh by direct communication with the patients of Sylhet Diabetes Hospital under the doctor supervision. Dataset had 520 samples and 16 features. Finally they got the accuracy of 99% from a Random Forest algorithm.

In [10], the author analyzed the dataset with several ML algorithms. Author used a public dataset from UCI machine learning repository where the dataset had 520 instances and 16 attributes. 10-fold cross validation used for result analysis and Random Forest gave the best accuracy of 96.9%.

In [11], the authors implemented three machine learning models named Logistic Regression, Naive Bayes and Decision Tree with Pima Indians Diabetes Dataset from UCI machine learning repository. They compared the results by accuracy and the Decision Tree showed the best accuracy of 77.9%.

3. Methodology

How the research is going on, all processes are shown in Figure 1. All the steps will be described one by one below.

3.1 Input Data

Most popular and commonly used Pima Indians Diabetes Dataset applied in this research. The author took that dataset from famous repository kaggle [17]. The dataset has 768 instances and 9 columns, attributes discretion included below in Table 1. Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age, and Outcome are the names of the 9 features. The target variable name is Outcome, it contains 1 or 0 means yes or no which represents whether the person has diabetes or not.

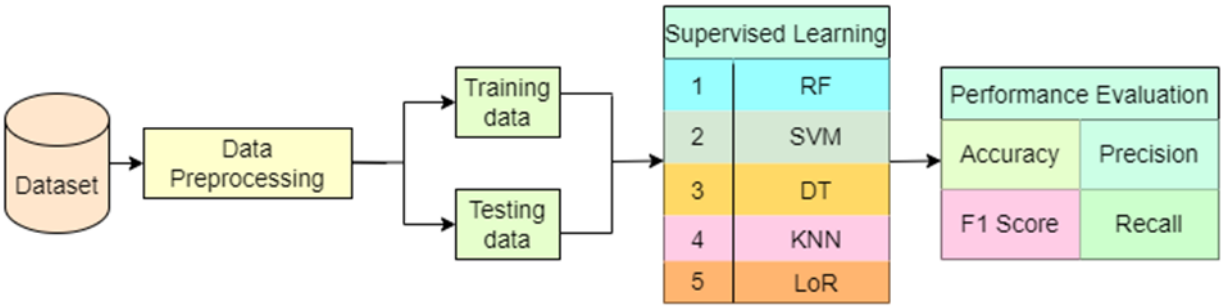

Figure 1: Methodological overview of the study

3.2 Data preprocessing

Any analysis of data initially starts with data preprocessing. There are so many ways to conduct data preprocessing. The author checked the null value first and the dataset has no null values, Random Forest, one of the supervised machine learning algorithms utilized in this work is unable to handle null values [18]. Data types are also checked due to categorical and numerical purposes. Data preprocessing is important because after the proper preprocessing, processed data will increase the model accuracy.

3.3 Split Data

In this study the dataset was divided into two subsets like training data and testing data. The 80:20 training to testing ratio signifies that 80% of the data is used to train the models and 20% is utilized to test them in this study. Before that the author separated the x data and y data, x data contains 8 columns out of 9 columns while y data contains only 1 column out of 9 columns. Independent variables indicate the x data and y data indicate the outcome or target or dependent variable. There are several ratios for splitting the dataset but which ratio the author used in this research gave the better result.

3.4 Correlation Matrix

How much related one variable to another variable and strength, direction of variables shows in correlation matrix. Feature selection process also needs to analyze the correlation matrix because highly correlated features sometimes negatively impact to reduce the performance for some models. So from the correlation matrix easily identify these types of features and delete or combine them to increase the performance. Dimensionality reduction, data cleaning, detecting collinearity, model interpretability are the things that can also be done with the help of correlation

matrix. -1 to 1 is the range of values, perfect positive correlation is denoted by a 1, perfect negative correlation by a -1, and no correlation is shown by a 0. The correlation between the dataset's various factors was displayed in Figure 2.

Table 1: Features description of dataset

9 Features	Description
Pregnancies	To demonstrate how many pregnancies there are
Glucose	To display the blood glucose level
Blood Pressure	To convey the measurement of blood pressure
Skin Thickness	To reflect the skin's thickness
Insulin	To display the blood's insulin level
BMI	To represent the Body mass index
Diabetes Pedigree Function	To express the percentage of diabetes
Age	To convey the age
Outcome	To articulate the ultimate outcome 1 indicates yes, and 0 indicates no

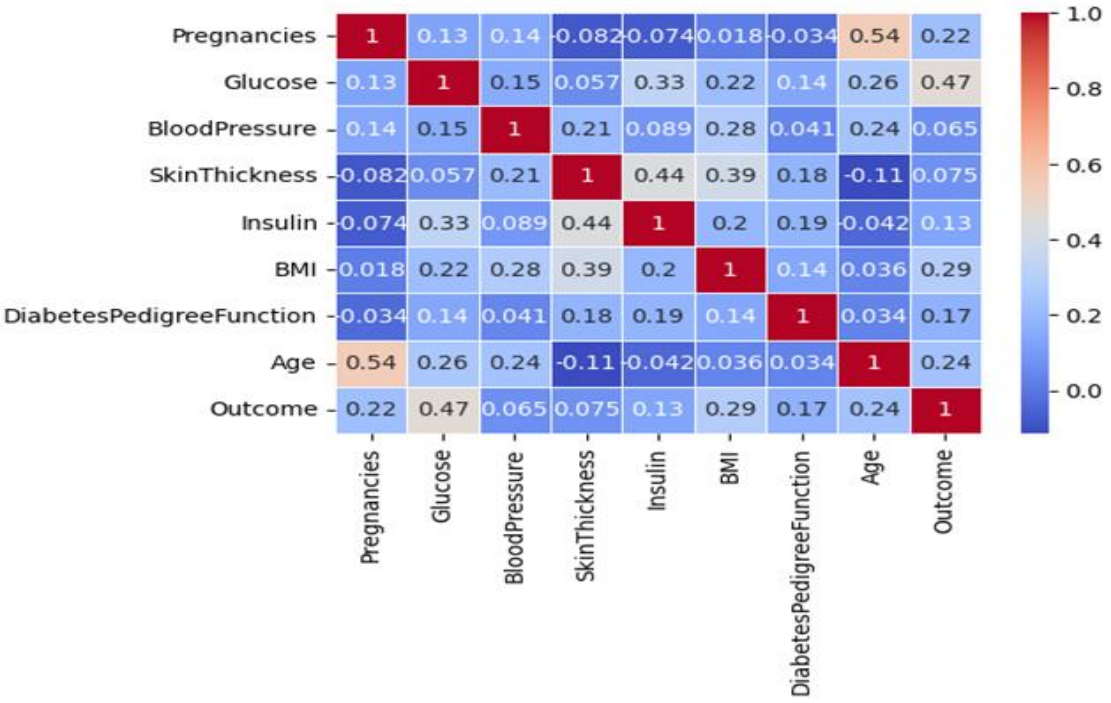

Figure 2: Correlation matrix between attributes of dataset

3.5 Applied ML Models

Random forest (RF): RF is one of the most popular and commonly used supervised learning techniques. It is basically a combination of decision trees. Classification and regression both problems can be solved by this algorithm. It is also widely used for ensemble learning. RF gives better accuracy when the number of decision trees is higher. Using the RF we can overcome the overfitting problems and get high accuracy [12], [13]. In this research the author used estimators equal to 100 of RF, and random state is equal to 1.

Support Vector Machine (SVM): The supervised learning method SVM is used to address regression and classification issues. Making a hyperplane from a proper or optimal decision boundary is the aim of support vector machines (SVM). The SVM is divided into classes from the n-dimensional space with the help of a hyperplane and the different classes hold the support vectors separately [13]. The author used probability equal to true and random state is equal to 1 in SVM for this research.

Decision tree (DT): It is called a decision tree because it has a structure like a tree. DT is also a commonly used supervised learning algorithm which allows both classification and regression. Each leaf node in the tree represents the conclusion, while internal nodes reflect a dataset's features and branches the decision rules. The decision tree has two nodes: the decision node and the leaf node. Multiple branches make up the decision node, which is utilized to make decisions. These decision nodes result in leaf nodes [14]. It has no additional branches. In the DT the author used a random state equal to 1.

K-Nearest Neighbors (KNN): One of the most straightforward machine learning algorithms that facilitates supervised learning is KNN. Classification and regression for both problems KNN will be used. It is also known as a non-parametric and lazy learner algorithm. KNN stores the new data with the most similar existing category after primarily identifying similarities between the new case and the available case [15].

Logistic Regression (LoR): LoR is also the supervised machine learning algorithm which is known as binary classifier. It is used to predict categorical data. We can use it for both classification and regression problems. LoR gives the output either 0 or 1, yes or no, true or false depending on independent and dependent variables [14], [16]. In this paper the author used a maximum of 200 iterations in LoR.

4. Result analysis and Discussion

Confusion matrix: Using the confusion matrix we can easily identify the accuracy, precision, recall and f1-score by calculating TP, TN, FP and FN [3]. An explanation of the confused matrix displayed in Table 2.

Table 2: Confusion matrix

	Predicted Negative	*Predicted Positive*
Actual Negative	TN (True Negatives) Correctly predicted negative cases.	FP (False Positives) Incorrectly predicted positive cases.
Actual Positive	FN (False Negatives) Incorrectly predicted negative cases.	TP (True Positives) Correctly predicted positive cases.

The equations of calculating accuracy, precision, recall and f1-score are given below in Eqs. 1, 2, 3 and 4, respectively [3]. Table 3 shows the percentages of each machine learning model's performance for different parameters.

Accuracy = (TP + TN) / (TP + TN + FP + FN) (1)
Precision = TP / (TP + FP) (2)
Recall = TP / (TP + FN) (3)
F1-Score = 2 * (precision * recall) / (precision + recall) (4)

The following provides descriptions of finding accuracy, precision, recall, and f1-score; Figure 3 displays the values of TP, TN, FP, and FN.
Accuracy: The model's overall correctness.
Precision: The accuracy of forecasts that are positive.
Sensitivity (Recall): The capacity to identify every incident of positivity.
The harmonic mean of recall and precision is the F1-Score.

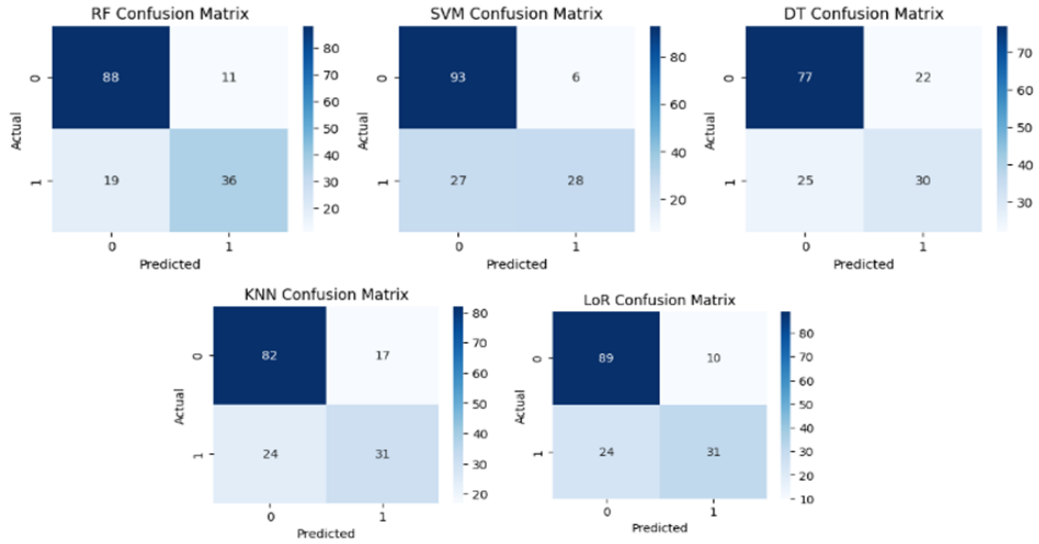

Figure 3: Confusion matrix of all applied ML models

The author assesses the prediction performance of several ML models by displaying the percentages of accuracy, precision, recall, and f1-score in Table 3, which is displayed below. Random Forest, SVM, Decision tree, K-Nearest Neighbors and Logistic Regression consecutively gave the accuracy of 81%, 79%, 69%, 73% and 78% while SVM gave the highest precision of 82%. Random Forest also gave the highest recall of 65% and f1-score of 71%.

Table 3: Performance comparison of all applied ML models

Algorithms	*Accuracy*	*Precision*	*Recall*	*F1 Score*
Random Forest (RF)	0.81	0.77	0.65	0.71
SVM	0.79	0.82	0.51	0.63
Decision Tree (DT)	0.69	0.58	0.55	0.56
K-Nearest Neighbors (KNN)	0.73	0.65	0.56	0.60
Logistic Regression (LoR)	0.78	0.76	0.56	0.65

Here you can see the comparison of different ML models performance in Figure 4. Random Forest is ahead of the other ML models.

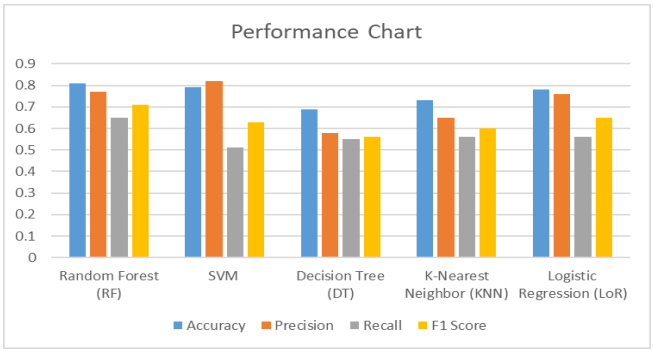

Figure 4: Performance comparison of all applied ML models

ROC curve: Performance measurement across thresholds, comparison of models, balanced performance metrics these are key points of ROC curve. By showing the ROC curve these things are easily visualized. The y-axis of the ROC curve represents the proportion of actual positives correctly identified, which means that true positive rate (TPR). The x-axis of the ROC curve indicates the proportion of actual negatives incorrectly identified as positives, which means false positive rate (FPR). Figure 5 shows the ROC curve of different ML models of this study. There are two types of varieties in supervised machine learning such as classification and regression, confusion matrix basically measure the classification enabled algorithms. Regression categories have different evaluation techniques. Categorical or discrete output instead of numerical or continuous we got from classification enables model and different prediction based analysis mostly done by categorical type datasets.

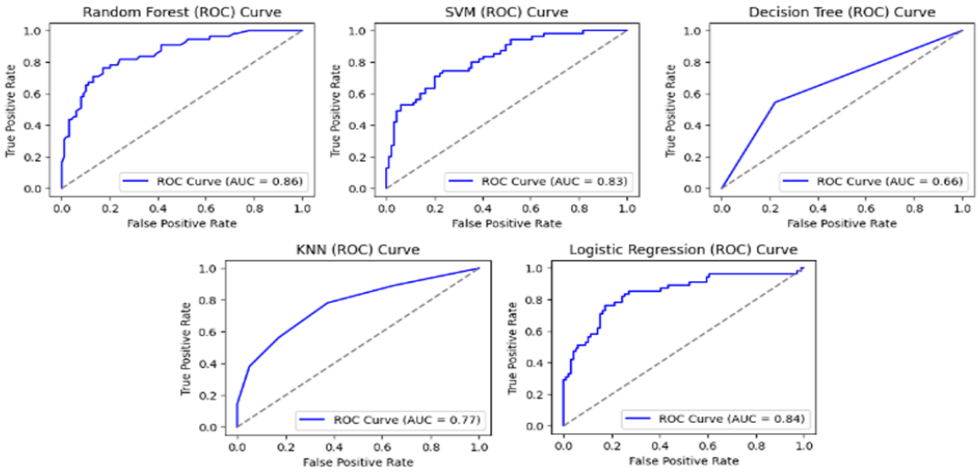

Figure 5: ROC curve of all applied ML models

From Figure 6 easily we can see that the best performing algorithm is Random Forest because it has the highest accuracy of 81% compared to other ML models.

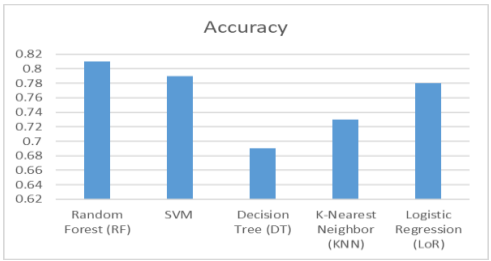

Figure 6: Accuracy comparison of all applied ML models

Finally the author concludes the result of this study by announcing Random Forest is the best machine learning algorithm for this research due to its high accuracy. Maybe this accuracy level is not up to the mark but through this research people can get the initial idea of some supervised machine learning algorithms based on their performance analysis. Hopefully in future the author will work to increase the performance level with some feature engineering.

5. Conclusion

Based on the above discussion, the accuracy of different ML models clearly can see that consecutively Random Forest, SVM, Decision Tree, K-Nearest Neighbor and Logistic Regressions are 81%, 79%, 69%, 73% and 78%. With an accuracy of 81%, Random Forest is the model that performs the best out of them all. Hope this study will help people, especially in the healthcare system. The continuous improvement of ML models and analysis with different datasets is a remarkable work for the society. Using this type of experiment has a good role in medical diagnosis for detecting disease early so that people take necessary actions to prevent these types of diseases. The limitations of this research, author did not include more feature engineering techniques in this paper. In the evaluating process there are more techniques available for testing models accurately, these also did not apply in this research. In the future the author will work with some advanced feature selection process such as PCA, IG, SHAP to identify the best suitable features for implementing the models to get more accuracy. Also work with useful techniques like cross validation for evaluating the models properly. Eventually the author can say that some other models like deep learning models with the collaboration of supervised learning can be done in future for finding the more actual results.

References

[1] S. Singh and V. Vazirani, “Classification vs clustering: Ways for diabetes detection,” in 2022 IEEE 7th International conference for Convergence in Technology (I2CT), IEEE, 2022, pp. 1–8.

[2] S. S. Reddy, N. Sethi, and R. Rajender, "A comprehensive analysis of machine learning techniques for incessant prediction of diabetes mellitus," Int. J. Grid Distrib. Comput., vol. 13, no. 1, pp. 1–22, 2020.
[3] H. A. Abdelhafez and A. A. Amer, "Machine Learning Techniques for Diabetes Prediction: A Comparative Analysis," J. Appl. Data Sci., vol. 5, no. 2, pp. 792–807, 2024.
[4] M. Emon, M. Keya, M. Kaiser, M. Islam, T. Tanha, and M. Zulfiker, Primary Stage of Diabetes Prediction using Machine Learning Approaches. 2021. doi: 10.1109/ICAIS50930.2021.9395968.
[5] G. Battineni, G. G. Sagaro, C. Nalini, F. Amenta, and S. K. Tayebati, "Comparative machine-learning approach: A follow-up study on type 2 diabetes predictions by cross-validation methods," Machines, vol. 7, no. 4, p. 74, 2019.
[6] I. J. Kakoly, M. R. Hoque, and N. Hasan, "Data-driven diabetes risk factor prediction using machine learning algorithms with feature selection technique," Sustainability, vol. 15, no. 6, p. 4930, 2023.
[7] A. Ram and H. Vishwakarma, "Diabetes prediction using machine learning and data mining methods," in IOP Conference Series: Materials Science and Engineering, IOP Publishing, 2021, p. 012135.
[8] K. Hirnak, N. Chaudhari, A. Singh, and D. Patil, "Early prediction model for type-2 diabetes based on lifestyle," in ITM Web of Conferences, EDP Sciences, 2020, p. 03053.
[9] V. Vakil, S. Pachchigar, C. Chavda, and S. Soni, "Explainable predictions of different machine learning algorithms used to predict Early Stage diabetes," ArXiv Prepr. ArXiv211109939, 2021.
[10] K. Kapoor, "How can machine learning be used to predict diabetes?".
[11] M. Kavitha and S. Subbaiah, "Implementing classification algorithms for predicting chronic diabetes diseases," Int J Eng Adv Technol, vol. 8, no. 6S3, pp. 1748–1751, 2019.
[12] H. El-Sofany, S. A. El-Seoud, O. H. Karam, Y. M. Abd El-Latif, and I. A. Taj-Eddin, "A Proposed Technique Using Machine Learning for the Prediction of Diabetes Disease through a Mobile App," Int. J. Intell. Syst., vol. 2024, no. 1, p. 6688934, 2024.
[13] A. M. Qadri, A. Raza, K. Munir, and M. S. Almutairi, "Effective feature engineering technique for heart disease prediction with machine learning," IEEE Access, vol. 11, pp. 56214–56224, 2023.
[14] S. Sanyal, D. Das, S. K. Biswas, M. Chakraborty, and B. Purkayastha, "Heart disease prediction using classification models," in 2022 3rd International Conference for Emerging Technology (INCET), IEEE, 2022, pp. 1–6.
[15] S. Bilgaiyan, T. I. Ayon, A. A. Khan, F. T. Johora, M. Parvin, and M. J. Alam, "Heart disease prediction using machine learning," in 2023 International Conference on Computer Communication and Informatics (ICCCI), IEEE, 2023, pp. 1–6.
[16] S. Yadav, A. Singh, V. Jadhav, and R. Jadhav, "Heart Disease Prediction Using Machine Learning," vol. Volume 9, pp. 761–765, Jul. 2023.
[17] "Diabetes Dataset - Pima Indians." Accessed: Jun. 28, 2024. [Online]. Available: https://www.kaggle.com/datasets/nancyalaswad90/review
[18] A. K. Dass, "Comparison of heart disease prediction using different machine learning algorithms," 2023.

Author Profile

Md. Jamaner Rahaman received the B.S. and M.S. degrees in Computer Science and Engineering from Daffodil International University in 2016 and 2018, respectively. Earlier he worked as a Research Associate in DIU and worked as a Lecturer in UCASM at the dept. of CSE. He is now working as a Lecturer since 01 February 2021 at the dept. of CSE in Leading University, Sylhet, Bangladesh.