INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

A Dataset-Specific Machine Learning Study for Predicting Diabetes (Type-2) in a Developing Country Context

Md Rakibul Haque[1], **Ibraheem Alharbi**[2]*

1 MIS Dept, University of Dhaka, Bangladesh
2 MIS Dept, College of Business, University of Jeddah, Kingdom of Saudi Arabia (KSA)

OPEN ACCESS

Received: 04-06-2022
Accepted: 17-08-2022
Published: 14-10-2022

Citation: Haque MR, Alharbi I (2022) A Dataset-Specific Machine Learning Study for Predicting Diabetes (Type-2) in a Developing Country Context. Indian Journal of Science and Technology 15(38): 1932-1940. https://doi.org/10.17485/IJST/v15i38.1183

* **Corresponding author**.

imalharbi@uj.edu.sa

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: Diabetes become more prevalent across the globe, understanding their sources and causes are more important than ever. This study uses machine learning techniques to efficiently detect Diabetic patients from many features. **Methods:** The purpose of this paper is to conduct a dataset-specific machine learning study for predicting diabetes in Bangladesh. Classification is used with 18 features including demographic characteristics, family history, dieting habit, clinical features, physical activities, and life quality. Five different classifiers are used. **Findings:** Based on using five different classifiers, results suggest that the Logistic Regression performed the best in predicting diabetes for this dataset. The accuracy of the logistic regression classifier exceeds 83.8%. **Novelty:** Unlike other studies, the authors combine eating habits with demographic and health features to enhance the performance of the classifiers. The result suggests that while addition of factors or features related to eating habits and lifestyle can increase the accuracy of prediction, the inclusion of more clinical features is more important to increase the accuracy. The authors believe that this finding is significant in the context of developing countries like Bangladesh considering the limited health-resource available as well as the fact of fast-changing of eating habits and lifestyle.

Keywords: Machine Learning; chronic disease; classification; logistic regression; and diabetes

1 Introduction

In 2021, International Diabetes Federation (IDF) reported that there are about 537 million diabetic patients globally where more than 4 in 5 or 81% of the patients reside in low-income and middle-income countries [(1)]. IDF also reported that one in 10 people worldwide is affected by diabetes and the global health expenditure has been increased by 316% in the last 15 years. Along with the fact that the number of diabetic patients has steadily increased over time, the number of deaths from diabetes has also increased. IDF recorded 6.7 million deaths worldwide caused by diabetes in 2021. Besides, the number of diabetic patients and diabetes-related healthcare and financial burden have raised sharply worldwide [(2)]. Diabetes-led chronic diseases have caused a financial burden in

every country[2].

The prevalence of diabetes also leads to the prevalence of other chronic diseases, and the consequently causing life-long sufferings, a decrease in the quality of life, and reduced life expectancy. Despite various factors like age, diet plan, level of physical activities, regular exercise, body mass index (BMI), lifestyle, obesity as well as genetic factors are known to contribute to the above-mentioned chronic diseases, one in every two people are unaware of the fact that they suffer from diabetes in general[1]. According to physicians and healthcare researchers, regular screening can delay the onset of diabetes. Therefore, given a wide range of factors that contribute to chronic diseases, predicting the likelihood of diabetes based on these factors is significant.

As the cause for the pathogenesis of diabetes is not known, researchers are continuously working to develop the most reliable and effective techniques for predicting diabetes based on various types of data including medical administrative data, genetic, and socio-economic data[3,4]. The use of advanced tools such as machine learning can predict the likelihood of diabetes by allowing us to learn and generalize from data[5]. Machine learning entails training based on data and allows learning about the complex relationships among various factors. It is widely used in the field of diabetes study for extracting knowledge and predicting the probability of diabetes and its related complications.

Unlike conventional statistical methods available for prediction, machine learning models use inductive inference for understanding how the risk factors interact with each other. The sophistication of machine learning allows us to identify trivial and nontrivial relationships among various factors from datasets,[6] which can then be interpreted to make logical predictions. Additionally, unlike the conventional statistical models, which require a time-consuming process of assuming multiple relationships between specific variables and identifying the predictor variable first, the machine learning models can select the predicting variables automatically and handle a large number of predicting variables[7]. Since the early diagnosis of the likelihood of diabetes is useful for controlling diabetes, many machine learning models have evolved in the last decade and shown promising results by analyzing many large and multi-dimensional datasets[8].

However, it is very challenging to increase prediction accuracy and precision when applying machine learning models. Especially, when the accuracy of machine learning significantly differs because of different datasets consisting of different types of features[9,10]. For instance, in a comparatively recent study, eight features are used from the Pima Indians Diabetes Database (PIDD) datasets and attainted an accuracy of up to 77.21%[11]. In the same study, they also used another dataset (Luzhou dataset) consisting of 14 features and achieved an accuracy of up to 80.84%. Also, different results were reported when the authors used three different datasets in the same study for cancer prediction[12]. Similarly, another study used five datasets containing same 54 features but collected in different periods of time and reported different levels of accuracy for each dataset when applied machine learning models to predict diabetes[13]. In this way, the machine learning results were compared using different datasets in different studies[14].

Therefore, the objective of this study is to conduct a dataset-specific machine learning study for predicting diabetes in the Bangladesh context. To accomplish this study, a dataset containing of 18 features was used as well as different machine learning models were used for performance comparison. Considering the study context, Bangladesh, which is a fast-developing country like many other developing countries in South and South-east Asia, the features related to demographic, eating habit and lifestyle were particularly added to the dataset in addition to the features related to family history, clinical diagnosis, and physical activities (Table 1). According to studies, demographic features[15], eating habit and life quality[16] are significantly linked with diabetes prevalence. These three factors are also changing fast in Bangladesh because of its fast-economic growth. Hence, the authors assume that the addition of the features related to these factors into the dataset would increase the accuracy of machine learning prediction.

This study data was collected from Bangladesh where the prevalence of diabetic patients is high. It is reported that there are 7.1 million diabetic patients across Bangladesh and the number is about to double by 2025[17]. It is also reported that there is also an equal number of diabetic patients in Bangladesh who are undetected. This study is significant because until now no similar study has been conducted with the above-mentioned study objective in the context of Bangladesh where the number of diabetic patients is growing fast. Besides, in Bangladesh where limited healthcare resources are unable to treat many diabetic patients, this study would help in prevention of diabetes Type 2 through early diagnosis.

A review of existing literature suggests that there are primarily three different types of machine learning techniques used for predictive research: 1) supervised learning which involves extracting information from labeled training data, 2) unsupervised learning which involved extracting information from unlabeled data. Additionally, reinforcement learning entails interaction between the machine learning algorithm and a dynamic environment[18]. While each has its strength(s), the supervised learning technique uses a target function that extracts information based on pre-selected features. On the other hand, the unsupervised learning technique finds the hidden relationships between unlabeled data, and the reinforcement learning extracts information through trials and errors when there is no prior knowledge regarding a new environment. This study has used the supervised

method as the dataset used in this study was consist of the features that the authors assume are related to diabetes onset based on different previous studies.

Researchers in the field increasingly use different machine learning toolkits, techniques, and models for predicting the possibility of many diseases with up to more than 95% of accuracy [9]. However, to extract more useful information, the methods and parameters should be selected reasonably. Besides, it is important to conduct the data pre-processing before applying machine learning models for better accuracy in prediction [19]. The most common types of data used for predicting diabetes are socio-economic data, vital signs, and diagnostic measurements. Until today, numerous studies have been conducted to predict diabetes using different machine learning models using a large number of factors associated with having diabetes. For example, recent study highlighted in a study by Meng et al. [13] that the demographic factors (e.g. age, education, income, and gender), family history, body measurement (e.g. BMI), lifestyle, daily activities, dieting habit are all significant predictive variables of diabetes [20–22].

The commonly used predictive models are Naïve Bayes (NB), Logistic Regression (LR), Random Forests (RF), K-nearest neighbor (k-NN), Decision Tree (DT), Support Vector Machine (SVM), and Artificial neural networks (ANN) [23]. However, interestingly, the best performing technique in one dataset might easily yield low-performance accuracy in another dataset [18]. Hence, it is important to compare the performance of the machine learning models for different datasets that differ based on the features, source, and types used.

The examples of machine learning models used in supervised learning are Support Vector Machines (SVM), Decision Trees (DT), and k-Nearest Neighbors (k-NN). The following sub-sections briefly discuss the five most popular and widely used predictive models.

1.1 Decision tree

Decision Tree is a powerful model often used by machine learning researchers. It makes decisions based on a flowchart like a tree where the tree starts with a single node containing the training samples. We used the J48 decision tree (also known as C4.5) that uses a top-down method that selects an attribute as a root node to generate a tree branch for every possible attribute and classify it into various subsets that are shown as connected to a root node [11].

In the Decision Tree model, the information of each variable is extracted after calculating the entropy of a dataset. Hence, if we consider two classes to be P and N, then the equations used in decisions tree are like:

$$Entropy = -\frac{P}{P+N}log_2\left(\frac{P}{P+N}\right) - \frac{N}{P+N}log_2\left(\frac{N}{P+N}\right)$$

And then we use:

$$IG(Y) = E(Y) - E(Y|X)$$

1.2 Random Forest

This model was proposed by Breiman that uses decision tree based on the Bagging technique [11]. It generates many decision trees to obtain more accuracy in prediction. The main difference between Decision Tree and Random Forest is that the latter finds the root node and split the feature nodes randomly. When predicting a variable, Random Forest generates separate classification results for each tree. A voting technique is then used to get the final output, which becomes the biggest number in the taxonomy.

1.3 Support Vector Machine

Support Vector Machine (SVM) is one of the most commonly used models and is considered as one of the state-of-the-art tools in machine learning. It is a supervised learning method that uses a training dataset to generate input-output mapping functions. The SVM algorithm uses both the classification and regression techniques to create a line in the graph that separates the data into classes, in which, data points with the same characteristics are grouped in the same class. Hyper-planes are constructed based on a given dataset which is considered as a p-dimensional vector that can be separated by a maximum of p-1 planes. The purpose of hyper-planes is to maximize the distance between two classes.

1.4 Logistic Regression (LR)

Logistic Regression (LR) model can predict the probability of the dependent variable which is a very popular way to fit models using categorical data. Logistic Regression uses a Sigmoid function instead of a Linear function where the values are restricted

to either 0 or 1 intervals.

The formula of logistic regression is:

$$P = \frac{1}{1 + e^{-(a + b_1X_1 + b_2X_2 + \ldots + b_nX_n)}}$$

1.5 K-Nearest Neighbor

K-Nearest Neighbor (KNN) is a very simple model, yet it gives very good results when dealing with regression and classification problems[24]. It is used in the supervised learning domain to find patterns. When training data is fed into this model as input, it finds out the classes, to which the new unlabeled object belongs. In this process, a 'K' is decided where k is the number of neighbors to be measured by a distance function. The majority voting is used in this case. This gives more accurate prediction.

2 Methodology

This study started with selecting and surveying 750 people. After cases with missing data were removed, data for a total of 738 people were usable. The average age of the participants was 44 years, with the highest age of 78 years old and the lowest age of 19 years old. Therefore, this study is concerned only with "adult-onset Diabetes Type-2 DM", as opposed to "Diabetes Type-1" which afflicts only children as an auto-immune disease. There were 256 participants (~34.6% from 738 participants) with diabetes and 482 non-diabetic participants.

The data were collected from both the rural and urban areas in Bangladesh. Table 1 shows the predicting factors and their types used in this study. In total, data related to 18 factors were used in the dataset. The supervised learning method was used for this study that required labeling the training data. Besides the literature review and WHO guidelines, two individual doctors were consulted for selecting the factors for questionnaire development and training data based on predicting factors. The questionnaire included a question related to the 18 factors as in Table 1. Out of 750 surveys, only 12 surveys were excluded because of incomplete or missing information.

Table 1. The types of data and factors used for training the model

Data Types	Factors (variables)
Demographic features	1. Area 2. Age 3. Gender 4. Marital Status 5. Education level
Family history	6. Family History of Diabetes
Eating habit (Weekly)	07. Eating Meat 08. Eating Fish 09. Eating Vegetables 10. Eating Fruits 11. Soft Drinks
Clinical diagnosis	12. Number of Pregnancies 13. BMI 14. Systolic Blood Pressure 15. Diastolic Blood Pressure 16. Glucose (Random)
Physical activities	17. Physical Activity
Lifestyle	18. Psychological stress

We employed a framework (Figure 1) that used five widely used machine learning models (i.e., Decision Tree (DT), Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), and K-nearest neighbor regression (K-NN) as described in the literature section above. For optimization, we tuned the parameters (Table 2) of the models accordingly. These classification models are known as widely used choices for classification tasks[25]. These models create model patterns to predict whether a patient has diabetes or not based on the labeled data provided. The dataset used was primary data which were classified into two classes: diabetic and non-diabetic.

We skipped the feature selection step of machine learning as the features were already selected based on previous studies that used the supervised learning technique. The machine learning algorithms were tested using the Orange package in the Anaconda Python environment. For model validation, we applied the k-fold cross-validation method to estimate the capability and skill of the model[11]. As we applied the cross-validation method, the input (Figure 1) data contained both the training and test data. For comparison results among the five used machine learning models, we calculated the accuracy, precision, recall, and F1 scores.

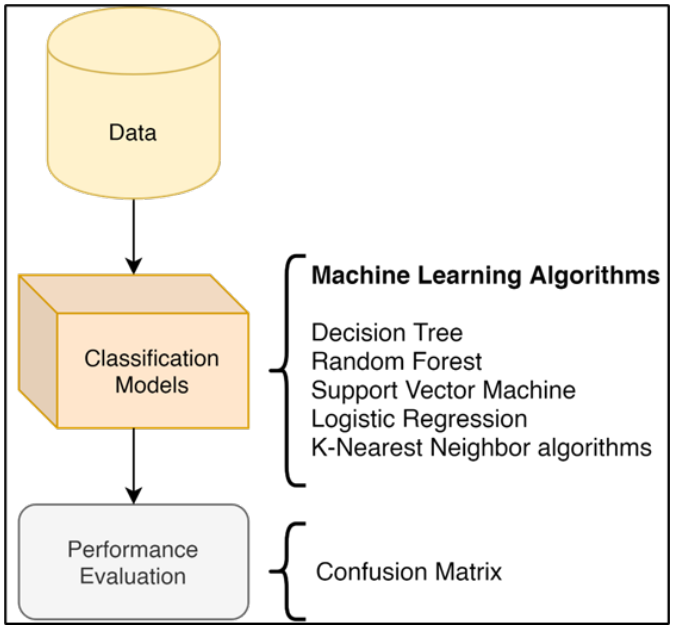

Fig 1. Steps used in machine learning

Table 2. Tuning Parameters used in the models

Model Names	Tuning Parameters
Decision Tree (DT)	Depth = 10
Random Forest (RF)	Estimators = 100
Support Vector Machine (SVM)	(N/A)
Logistic Regression (LR)	(N/A)
K-nearest neighbor regression (K-NN)	Neighbor = 5

3 Results and discussion

The supervised learning technique with five machine learning models was employed. Figure 2 shows that the five models were evaluated against the parameters called Accuracy, Precision, Recall, and F1 Score[24]. The Accuracy gives us information regarding how often the prediction was correct. It is the ratio of the number of correct predictions to the total predictions. The equation of Accuracy is Accuracy = TP+TN/TP+FP+FN+TN. Here, TP, TN, FP, and PN are known as True Positives, True Negatives, False Positives, and False Negatives respectively.

According to Table 3 below, TP means that the value of both the actual class and predicted class is Yes; and the TN means that the value of both the actual class and predicted class is No. Conversely, FP means that the value of the actual class is No, but the value of the predicted class is Yes. Likewise, FN means that the value of the actual class is Yes, but the value of the predicted class is No.

Table 3. Actual class and predicted class matrix

	Predicted class		
		Class = Yes	Class = No
Actual class	Class = Yes	True Positives	False Negatives
	Class = No	False Positives	True Negative

We also calculated the value of Precision which implies the ability of a model of correct positive prediction. It is the ratio of the number of correct positive predictions to the total positive predictions. The equation of Precision is Precision = TP/TP+FP. Additionally, we calculated the Recall (also known as sensitivity) which is the ratio of correct positive predictions to all predictions in actual class - Yes. The equation of Recall is Recall = TP/TP+FN.

Furthermore, we calculated the F1 Score which is simply the value of the weighted average of Precision and Recall. The value of the F1 Score is more useful as it takes into account both the Recall and Precision, especially when it is not easy to understand the performance of a model based on Accuracy. The equation of F1 Score is: F1 Score = 2*(Recall * Precision) / (Recall + Precision).

hows that the Logistic Regression model had the highest Accuracy (83.8%), Precision (86.0%), and F1 Score (70.0%) in comparison to other models. The Random Forest has yielded the second most Accuracy (82.4%), Precision (80.0%), and F1 Score (69.8%). On the other hand, the SVM has yielded the highest Recall (71.0%) but the lowest Accuracy (66.2%). Besides, K-nearest neighbour regression (K-NN) has shown the lowest F1 Score (54.9%). Overall, Logistic Regression showed the best result out of five models.

Table 4. Comparing the accuracy and precision of the performance of each machine learning technique.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
Logistic Regression	83.8	86	62	70
Random Forest	82.4	80	64	69.8
Decision Tree	73.7	59	72	68.4
KNN	71.0	57	58	54.9
SVM	66.2	55.4	71	62.2

Figure 2 shows that out of 18 features, the most important feature is blood glucose followed by age, systolic blood pressure and BMI. Figure 2 shows that if the feature glucose is removed from the data, then the accuracy of prediction decreases by about 30%. On the other hand, the absence of age, systolic blood pressure and BMI from the data decrease the accuracy of prediction by 11.2%, 8.9% and 8.4% respectively. It implies that clinical data like blood glucose are stronger features among all types of features (i.e., demographic features, family history, eating habit) added. Besides, Figure 3 (Y-axes is Glucose level and x-axes is BMI) shows that the accuracy by Logistic Regression (.80) can be more generalised than any other models used. It implies that if Logistic Regression is used in other future studies with similar dataset features used in this study, then that may also show more accuracy in comparison to other models as found in this study.

To discuss the results, we used 18 features to test a dataset-specific machine learning study the for prediction of diabetes. The main objective of this study was to conduct a dataset-specific machine learning study for predicting diabetes in the Bangladesh context and compare the prediction performance of different machine learning models in terms of their accuracy, precision, recall, and F1 Score. According to the authors' best knowledge, no studies have previously examined the Machine Leaning models with a dataset consisting of 18 particular features in Bangladesh for predicting the possibility of diabetes. The dataset contains data of six different types of factors as shown in Table 1. Based on our analysis, we can see that the Logistic Regression (LR) model is the best predicting model (Accuracy of 83.8%).

When comparing with other study, the highest accuracy in this study was 83.7%, which is more than the highest accuracy of 82.10% as shown by another recent study (Ahmad et al.)[10] in which the dataset that was used had different features. The main difference between the dataset used for this study and the dataset used by Ahmad et al. is that this study has used features related to eating habits and lifestyle. Hence, in comparison to their study, the result in this study shows more accuracy which is important contribution in the context of use of machine learning for disease prediction. Unlike previous study, the comparative higher accuracy reported in this study implies that future studies need to consider adding more features or factors related to eating habits and lifestyle to increase the accuracy of prediction.

However, although previous studies claimed that the non-clinical factors such as demographics, eating habits, and lifestyle in Table 1 influence diabetes prevalence significantly, the inclusion of these features in the dataset in this study has not greatly increased the accuracy. The accuracy level increases up to 88.27% when used clinical factors like fasting plasma glucose (FPG). It implies that selection of specific clinical data increases the accuracy of model prediction drastically. This finding is also supported by the result (Figure 3) in this study that shows if the feature glucose is removed from the data, then the accuracy of prediction decreases by about 30%. This indicates that the dataset consists of more clinical features such as HbA1c test and Glucose level[26] are more important than the dataset consisting of the features related to demographics, eating habits, and lifestyle for diabetes prediction using machine learning models. Therefore, despite having more accuracy in this study due to adding non-clinical factors or features such as eating habits and lifestyle, clinical data was found to be the biggest determinant in predicting diabetes than the features such as demographics, eating habits, and lifestyle except age, which is a significant finding that has not reported by previous studies.

Also, previous study[10] have shown different machine learning models (e.g., SVM and k-NN) as the best performer for prediction, we know that even the best model might show low performance in prediction with different datasets derived from

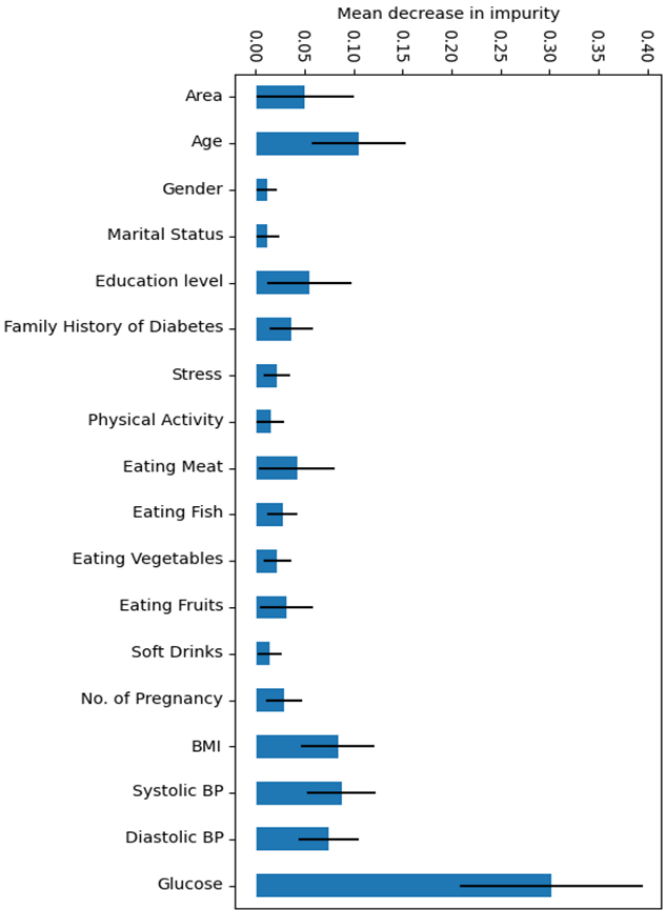

Fig 2. Comparing the most important features.

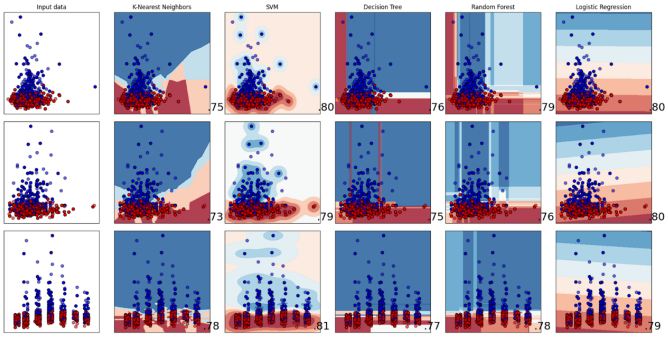

Fig 3. Projecting the decision boundary for diabetes prediction task by all five models, where Y-axes of all the plots are showing Glucose level. On the other hand, plots in 1st row uses BMI as x-axes, plots in 2nd row uses Systolic BP as x-axes, and plots in 3rd row uses Age as x-axes. The plots show training points in solid colors and testing points semi-transparent. Test accuracy is mentioned at the bottom right of each plot.

another context. In our study, Logistic Regression (LR) was found to be the best performer in predicting diabetes. However, as accuracy alone may not confirm good performance, in our study, Logistic Regression (LR) has also shown the highest values in terms of precision, recall, and F1 Score.

Even though the dataset used in this study can predict diabetes in most of the cases (as F1 score for LR is 70.0%), without the inclusion of more clinical features as well as the features related to demographic, eating habit, lifestyle, the machine learning models cannot significantly increase the accuracy. For instance, other three recent studies[27–29] used clinical features such as insulin level and skin thickness, and their study attained the accuracy levels of 90%, 88.6%, and 96% respectively. Hence, in comparison to previous studies, this study strongly encourages the inclusion of more clinical features in future studies in addition to the features related to eating habits and lifestyle.

4 Conclusions

In this study, we proposed a machine learning-based framework and applied the primary data in the local context of Bangladesh. The main objective of this study was to conduct a dataset-specific machine learning study for predicting diabetes in Bangladesh. The authors used five different machine learning models to compare the results. The result shows that Logistic Regression was the best (accuracy was 83.8%) in classifying patients with diabetes. Although the result is promising, the authors suggest for incorporating more features related to clinical diagnosis to increase the accuracy more than 95%.

Based on the findings, the major contribution of this study is to the adding of features related to the eating habit and lifestyle, which have increased the accuracy of prediction in comparison to previous studies as discussed above. In addition, the authors conclude that the inclusion of more clinical features is more important than the inclusion of the features related to demographic, eating habit, and lifestyle, which previous studies have not reported by comparing the results.

Finally, the authors suggest that future studies should also use more complex machine learning techniques such as ensemble method to optimize the dataset and ensure that prediction of diabetes using machine learning makes more sense and is acceptable to the medical personnel, users, and stakeholders. The authors believe that this finding is significant because accurate and timely prediction of chronic diseases, such as diabetes, is important in saving lives, improving the quality of life, and extending average life expectancy, especially in the context of developing countries like Bangladesh.

References

1) IDF. Diabetes facts & figures. Diabetes Atlas 10th Edition. 2021. Available from: https://diabetesatlas.org/.
2) Htay T, Soe K, Lopez-Perez A, Doan AH, Romagosa MA, Aung K. Mortality and Cardiovascular Disease in Type 1 and Type 2 Diabetes. *Current Cardiology Reports*. 2019;21(6):45–45. Available from: https://doi.org/10.1007/s11886-019-1133-9.
3) Panicacci S, Donati M, Profili F, Francesconi P, Fanucci L. Trading-Off Machine Learning Algorithms towards Data-Driven Administrative-Socio-Economic Population Health Management. *Computers*. 2021;10(1):4–4. Available from: https://doi.org/10.3390/computers10010004.
4) Awotunde IDJBO, Babatunde O. An Improved Hybridization in the Diagnosis of Diabetes Mellitus Using Selected Computational Intelligence. *Information and Communication Technology and Applications: Third International Conference ICTA 2020 Minna Nigeria*. 2020. Available from: https://doi.org/10.1007/978-3-030-69143-1_22.
5) Bosnyak Z, Zhou FL, Jimenez J, Berria R. Predictive Modeling of Hypoglycemia Risk with Basal Insulin Use in Type 2 Diabetes: Use of Machine Learning in the LIGHTNING Study. *Diabetes Therapy*. 2019;10(2):605–615. Available from: https://doi.org/10.1007/s13300-019-0567-9.
6) Saberioon M, Císař P, Labbé L, Souček P, Pelissier P, Kerneis T. Comparative Performance Analysis of Support Vector Machine, Random Forest, Logistic Regression and k-Nearest Neighbours in Rainbow Trout (Oncorhynchus Mykiss) Classification Using Image-Based Features. *Sensors*. 2018;18(4):1027–1027. Available from: https://doi.org/10.3390/s18041027.
7) Steele AJ, Denaxas SC, Shah AD, Hemingway H, Luscombe NM. Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease. *PLOS ONE*. 2018;13(8):e0202344–e0202344. Available from: https://doi.org/10.1371/journal.pone.0202344.
8) Chauhan T, Rawat S, Malik S, Singh P. Supervised and Unsupervised Machine Learning based Review on Diabetes Care. *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*. 2021;p. 581–585. Available from: https://doi.org/10.1109/ICACCS51430.2021.9442021.
9) Chaki J, Ganesh ST, Cidham SK, Theertan SA. Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review. *Journal of King Saud University - Computer and Information Sciences*. 2022;34(6):3204–3225. Available from: https://doi.org/10.1016/j.jksuci.2020.06.013.
10) Ahmad HF, Mukhtar H, Alaqail H, Seliaman M, Alhumam A. Investigating Health-Related Features and Their Impact on the Prediction of Diabetes Using Machine Learning. *Applied Sciences*. 2021;11(3):1173–1173. Available from: https://doi.org/10.3390/app11031173.
11) Zou Q, Qu K, Luo Y, Yin D, Ju Y, Tang H. Predicting Diabetes Mellitus With Machine Learning Techniques. *Frontiers in Genetics*. 2018;9:515–515. Available from: https://doi.org/10.3389/fgene.2018.00515.
12) Aalaei S, Shahraki H, Rowhanimanesh A, Eslami S. Feature selection using genetic algorithm for breast cancer diagnosis: experiment on three different datasets. *J Basic Med Sci*. 2016;19(5):476–482. Available from: https://doi.org/10.3389/fgene.2018.00515.
13) Meng XH, Huang YX, Rao DP, Zhang Q, Liu Q. Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. *The Kaohsiung Journal of Medical Sciences*. 2013;29(2):93–99. Available from: https://doi.org/10.1038/s41598-020-68771-z.
14) Rani AS, Jyothi S. Performance analysis of classification algorithms under different datasets. *3rd International Conference on Computing for Sustainable Global Development (INDIACom)*. 2016;p. 1584–1589.

15) Tung EL, Chin MH. Demographic Influences and Health Disparities in Adults with Diabetes. *Behavioral Diabetes*. 2020;p. 441–461. Available from: https://doi.org/10.1007/978-3-030-33286-0_28.
16) Singh D, Leavline EJ, Baig BS. Diabetes prediction using medical data. *J Comput Intell Bioinforma*. 2017;10(1):1–8. Available from: https://doi.org/10.37896/JXAT14.01/314405.
17) Mohiuddin AK. Diabetes Fact: Bangladesh Perspective. *Int J Diabetes Res*. 2019;2(1). Available from: https://doi.org/10.17554/j.issn.2414-2409.2019.02.12.
18) Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine Learning and Data Mining Methods in Diabetes Research. *Computational and Structural Biotechnology Journal*. 2017;15:104–116. Available from: https://doi.org/10.1016/j.csbj.2016.12.005.
19) Vijayan VV, Anjali C. Decision support systems for predicting diabetes mellitus-A Review. Global Conference on Communication Technologies (GCCT). 2015;p. 98–103. Available from: https://doi.org/10.1109/GCCT.2015.7342631.
20) Pinchevsky Y, Butkow N, Raal FJ, Chirwa T, Rothberg A. Demographic and Clinical Factors Associated with Development of Type 2 Diabetes: A Review of the Literature</p>. *International Journal of General Medicine*. 2020;13:121–129. Available from: https://doi.org/10.2147/IJGM.S226010.
21) Talukder A, Hossain MZ. Prevalence of Diabetes Mellitus and Its Associated Factors in Bangladesh: Application of Two-level Logistic Regression Model. *Scientific Reports*. 2020;10(1):1–7. Available from: https://doi.org/10.1038/s41598-020-66084-9.
22) Kyrou I, Tsigos C, Mavrogianni C, Cardon G, Van Stappen V, Latomme J, et al. Sociodemographic and lifestyle-related risk factors for identifying vulnerable groups for type 2 diabetes: a narrative review with emphasis on data from Europe. *BMC Endocrine Disorders*. 2020;20(S1):1–13. Available from: https://doi.org/10.1186/s12902-019-0463-3.
23) Dagliati A, Marini S, Sacchi L, Cogni G, Teliti M, Tibollo V, et al. Machine Learning Methods to Predict Diabetes Complications. *Journal of Diabetes Science and Technology*. 2018;12(2):295–302. Available from: https://doi.org/10.1177/1932296817706375.
24) Kaur H, Kumari V. Predictive modelling and analytics for diabetes using a machine learning approach. *Applied Computing and Informatics*. 2022;18(1/2):90–100. Available from: https://doi.org/10.1016/j.aci.2018.12.004.
25) Zheng T, Xie W, Xu L, He X, Zhang Y, You M, et al. A machine learning-based framework to identify type 2 diabetes through electronic health records. *International Journal of Medical Informatics*. 2017;97:120–127. Available from: https://doi.org/10.1016/j.ijmedinf.2016.09.014.
26) Cahn A, Shoshan A, Sagiv T, Yesharim R, Goshen R, Shalev V, et al. Prediction of progression from pre-diabetes to diabetes: Development and validation of a machine learning model. *Diabetes/Metabolism Research and Reviews*. 2020;36(2). Available from: https://doi.org/10.1002/dmrr.3252.
27) Khanam JJ, Foo SY. A comparison of machine learning algorithms for diabetes prediction. *ICT Express*. 2021;7(4):432–439. Available from: https://doi.org/10.1016/j.ijmedinf.2016.09.014.
28) Daanouni O, Cherradi B, Tmiri A. Diabetes Diseases Prediction Using Supervised Machine Learning and Neighbourhood Components Analysis. *Proceedings of the 3rd International Conference on Networking, Information Systems & Security*. 2020;p. 1–5. Available from: https://doi.org/10.1145/3386723.3387887.
29) Mujumdar A, Vaidehi V. Diabetes prediction using machine learning algorithms. *Procedia Comput Sci*. 2019;165:292–299. Available from: https://doi.org/10.13140/RG.2.2.21353.21603.