Diabetes Prediction Using Artificial Intelligence

M.M. Shahriar Amin   
*Department of CSE  
Brac University*Dhaka,Bangladesh

[m.m.shahriar.amin@g.bracu.ac.bd](mailto:m.m.shahriar.amin@g.bracu.ac.bd)

Partho Mark Gomes  
*Department of CSE   
Brac University*Dhaka,Bangladesh

parthomark@gmail.com  
 Faiza Tasneem *Department of CSE  
 Brac University* Dhaka,Bangladesh  
faizneem97@gmail.com

Towshik Anam Taj  
*Department of CSE  
Brac University*Dhaka,Bangladesh towshiktaj6@gmail.com

*Abstract*— Artificial Intelligence gained a significant position in healthcare services (HCS) due to its ability to improve the disease prediction in HCS. Artificial intelligence has already been worked in the HCS area. Recently, diabetes is a notable public chronic disease worldwide. It is growing rapidly because of bad lifestyles, taking more junk food and lake of health awareness. Therefore, there is a need of framework that can effectively track and monitor people’s diabetes and health condition within an application view. The literature problem is that they use skin column in the dataset which we drop because skin and Skin Thickness are correlated 1 and 1. So, we didn’t use the Skin column in our dataset. Also, we replace the zero mean by mean which is easy and fast. We also used RF model which gave 77% accuracy and our LR model gave 78% instead of 67% in the previous literature. In this work, six most important artificial intelligence classification techniques were considered for predicting diabetes. However, we use different evaluation criteria to investigate the performance of these classification techniques. In addition, performance measurement of the classification techniques was evaluated by applying the 10-fold cross validation method. The analysis results show that Logistics Regression model achieved highest performance than the other classifiers, obtaining the F1 measure of 0.78.

# *1. Introduction*

Diabetes Mellitus (DM) is defined as a group of metabolic diseases in which humans have towering blood sugar levels. Diabetes is a prolonged disease that happens when the body cannot efficiently use the insulin it generates. As a result, the disease increase the risk of malfunction of different organs, especially the eyes, kidneys, nerves, heart, and blood vessels .The deaths of around 1.6 million people were completely affected by diabetes in 2015 and 2.2 million deaths due to high blood glucose in 2012[1]. Diabetes Mellitus do not depend on the age, it can happen with people anytime. There are three types of diabetes [2]: i) Juvenile or childhood diabetes (type 1 diabetes), ii) Type 2 or adult diabetes iii) Gestational or type 3 diabetes. Gestational diabetes is hyperglycemia which occurs because of the change in hormones during pregnancy. Generally, type 1 diabetes happens due to the lack of insulin production and it is diagnosed in people of young age [2]. Type 2 is a very familiar form of diabetes, and it contains a huge volume of people from around the world. Type 2 mostly causes surplus body weight and physical disuse. Whatsoever, type 1 and type 2 diabetes cannot be cured properly. But early diagnosis and simple lifestyle can prevent it. Moreover, there are different new cases of diabetes arising from the developing countries [3] where in Bangladesh, the number of

diabetes affected people are increasing manifold. In last few decades, data has been elevated in a vast scale in diverse arenas including medical fields [4]. Artificial Intelligence is a discipline that aims to solve different important biomedical problems. The artificial intelligence-based classification techniques are the most operative methods for both real-life and scientific problems [5]. The use of these classification-based approaches in the diagnosis and cure of diseases can significantly decrease medical errors and human costs. Machine learning based classification techniques have prospective performance in prediction accuracy as compared to other algorithms for data classification. Data classification accuracy may vary conditionally on different machine learning techniques. Many of the researchers have been focused on diabetes from various perspectives of their works where most of the study discussed the classification techniques for diabetes prediction and its accuracy. The main aspect of this project is to obtain more efficient results and aim to be more successful afterwards. Therefore, the aim of this study is to evaluate the performance of different artificial intelligence classification techniques for the classification of diabetic affected or not. We use six classification techniques, Artificial Neural Network (ANN), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Logistics Regression (LR) and Naïve Bayes (NB). We explore the performance result of different techniques where the performance is evaluated by various standards, such as accuracy, precision, F-score and Recall. Moreover, the most accurate classification technique is donated for diagnosis of such disease with proposed unified framework.

# PROPOSED MODEL

We used ANN, SVM, LR, DT, RF and NB in our projects. Description of each model is given below

*2.1**Artificial Neural Network (ANN):*

Artificial neural network (ANN) is an important machine learning technique for biological research. In machine learning, ANN is a convenient computational model which works like biological neurons. Elementary structure of ANN is a collection of linked nodes. Moreover, these nodes help to perform as neurons in ANN, considering the nodes are connected by a link and each link has some weight. ANN mainly organized into three layers; i) Input layer (nodes can take input data), ii) Hidden layer or processing stage (processes the input data from input layer) and iii) output layer (results are sent from the hidden layer). In addition, the result of output layer in each node is called its activation or node value

* 1. *Support Vector Machine (SVM)*

Support vector machine (SVM) is a supervised learning algorithm which is based on linear classification. SVM work well for many health care problems and can solve both linear and non-linear problems. In order to solve the regression and classification problems efficiently SVM perform better than other classification techniques. Therefore, Vladimir Vapnik and Alexey Chervonenkis [6] introduced the support vector machine classification technique which is attempted to pass a linearly separable hyperplane to classify the dataset into two classes. Finally, the model can undoubtedly estimate the target groups (labels) for new cases.

* 1. *Logistics Regression (LR)*

Logistic Regression was mostly used in the biological research and applications in the early 20th century [7]. Logistic Regression (LR) is one of the most used machine learning algorithms that is used where the target variable is categorical. Recently, LR is a popular method for binary classification problems. Moreover, it presents a discrete binary product between 0 and 1. Logistic Regression computes the relationship between the feature variables by assessing probabilities (p) using underlying logistic function.

* 1. *Decision Tree (DT)*

Decision tree (DT) is one of the popular supervised learning-based classification algorithms in Machine Learning. DT can be used for both classification and regression problems. Moreover, DT is a classification technique which breaks a dataset into smaller subsets or a composite decision into a union of several easier decisions, at the meantime, the final solution with associated decision tree is incrementally developed [8].

* 1. *Random Forest (RF)*

Random Forest (RF) is a well-known supervised classification algorithm which is able to perform both regression and classification. RF has been first proposed by Leo Breiman [9]. In general, RF constructs several decision trees and combines them

together to acquire more accurate and efficient prediction. These techniques add an extra layer of randomness to bagging. Moreover, the random-forest algorithm fetches a subset of predictors randomly preferably at the node where the trees splits.

* 1. *Naive Bayes (NB)*

Naive Bayes classifier is a simple but most operative algorithm for the classification problems. Naive Bayes are statistical classifiers that works by making a hypothesis of conditional independence with the training datasets [10]. Henceforth, Naive Bayes classifier is the appropriate classification technique that

verdicts best solution for a dataset from a pool of different objects.

# *Diagram*

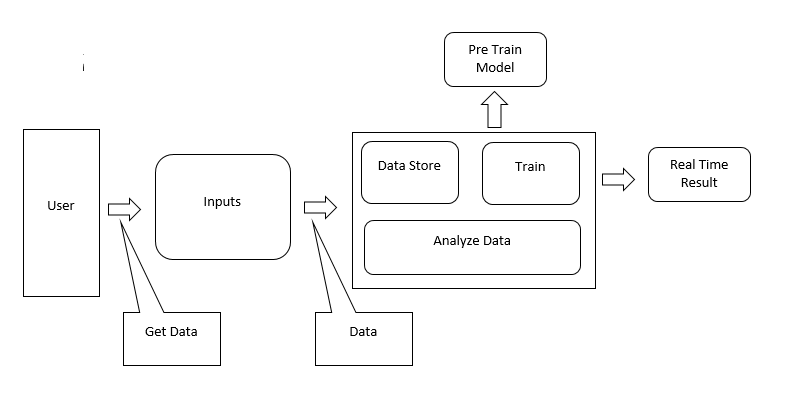


Figure 1

As we can see from the diagram, Inputs will be taken from the user.The inputs are meant to be large datasets here.Then we will process the data and make the dataset free from null and garbage values.Then we’ll split it into two portion and will train train the network using train dataset and meanwhile we’ll analyze the data as well.Ex, find standard deviation,compare between fields etc to completely uderstand the dataset and then finally will get our desired output using test dataset once training is done.Then we’ll measure and compute rate of error and other things.

# *exPERIMENTS*

We have used PIMA Indian Dataset from Kaggle.com. In this project, we used different classification algorithm to train and test our model. Performance of all the classification algorithms are assessed by different statistical measurement aspects such as accuracy, specificity, f1-score etc. These classification measurement factors are calculated by the terms: True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN). Here,

From the Prima Indian Data set, 268 true samples and 500 negative samples were taken into analysis. We split the diabetes data set into two parts where the training set contains 70% and the test set contains the remaining 30% of the data, where, training true: 188 (35.01%), training false: 349 (64.99%), test true: 80 (34.63%) and test false: 151 (65.37%). Moreover, the dataset was also checked to verify the correlated features in order to drop the redundant columns.

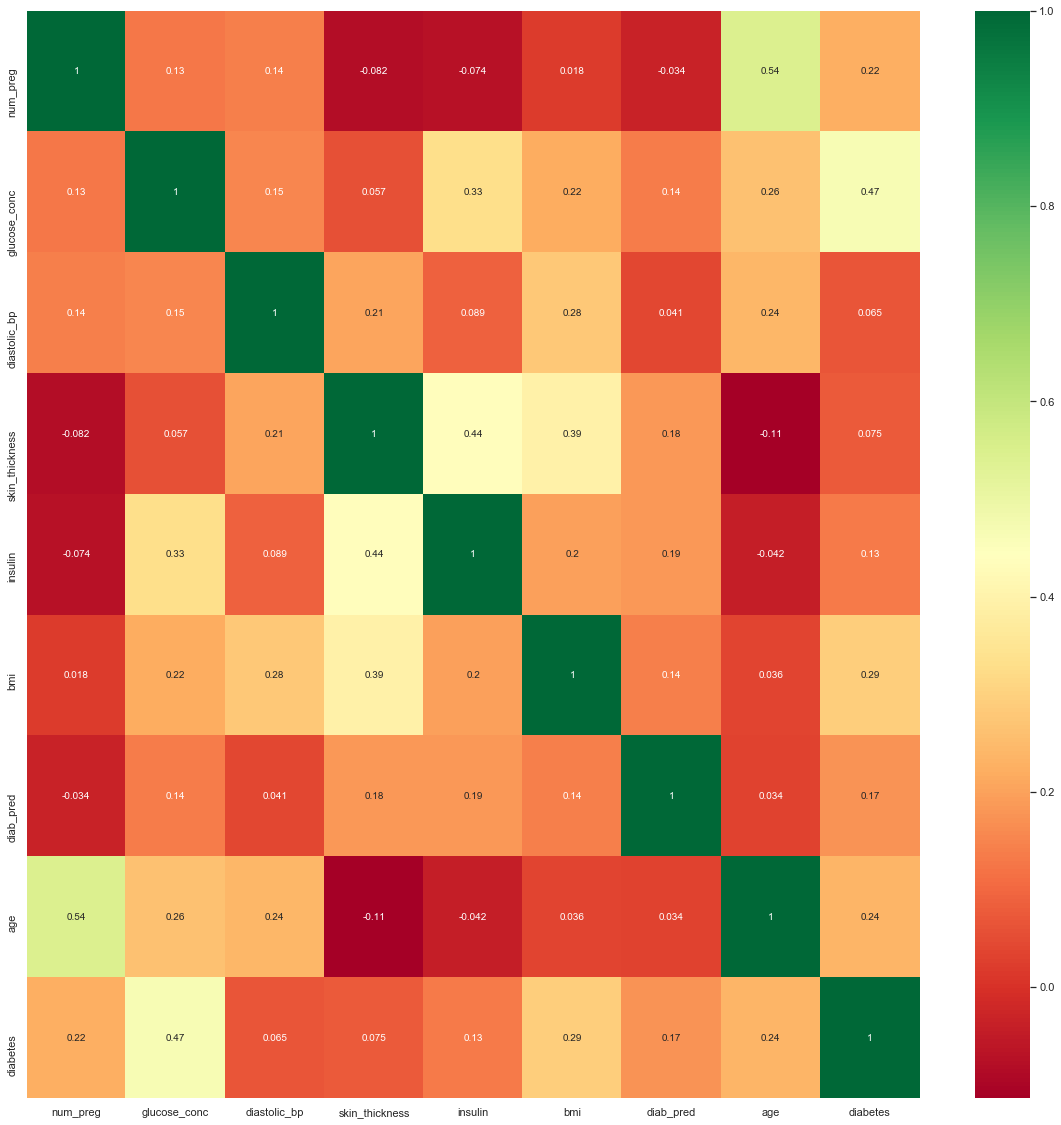


Figure 2. Red is most correlated, Blue least.

Input: data\_frame: pandas DataFrame

size: vertical and horizontal size of the plot

The K-fold cross validation approach is used to evaluate the performance of the prediction model. Predictions of all the machine learning classification algorithms ,Random Forest (RF) and Logistics Regression (LR) exhibits the highest performance and Decision Tree Model ( DT ) shows the lowest performance than the other 5 classification algorithms in terms of the four measurement factors: recall, precision and F-1 score measure in table 1.

Table 1. Classification performance measurements on the dataset

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Measurement**  **Techniques** | **N**B | RF | LR | SVM | ANN | DT |
| Accuracy | .76 | .77 | .78 | .77 | .77 | .70 |
| Precision | .74 | .75 | .77 | .75 | .75 | .67 |
| F-1 Score | .74 | .77 | .78 | .76 | .77 | .70 |
| Recall | .73 | .74 | .74 | .72 | .77 | .70 |

All the machine learning classifiers show the accuracy level of nearly 70%, which indicates that the performance of these techniques is well. F-1 measure indicates (NB, SVM, DT, LR and RF) that the five-classification techniques mostly predict accurate results. From the above discussion, it is important to know about the Receiver Operating Characteristics (ROC) curve, which is based on the true positive rate (TPR) and false positive rate (FPR) of these classification results. The ROC curve is presented in Figure 3.

A screenshot of a cell phone

Description automatically generated

we highlight the research directions and scope in relation to Health Care Services (HCS) and Bio-medical fields by machine learning classification techniques. Hence, disease prediction by machine learning classification algorithms should be improved. We describe the most popular Artificial Intelligence techniques and give the purpose of our project which is a unified framework for diabetes prediction that require further research in terms of machine learning based disease prediction.

##### *5. COnclusion*

The main contribution of this study is that we compare the performance of the six-machine learning classification techniques and evaluated their performance using the 10-fold validation technique.

In general, multiple machine learning classifiers should perform better than a single machine learning classifier. The experimental result show that the highest classification accuracy is 74% and highest F1 measure is 0.74. In addition, this application can classify the patients based on their diabetes level by collecting real time data from various Health Care Services, such as medical diagnose center, hospital, health tracking devices and sensors etc.

##### *6. References*

[1] American Diabetes Association. 2010. Diagnosis and classification of diabetes mellitus. Diabetes care. 33 Suppl 1, (Jan. 2010), S62-9. DOI= 10.2337/dc10-S062.

[2] Diabetes: 2017. http://www.who.int/news-room/factsheets/detail/diabetes. Accessed: 2018-06-08.

[3] Samant, P., and Agarwal, R. 2018. Machine learning techniques for medical diagnosis of diabetes using iris images. Computer Methods and Programs in Biomedicine. 157, (Apr. 2018), 121–128. DOI: <https://doi.org/10.1016/J.CMPB.2018.01.004>.

[4] Ahmed, M. R., Khatun, M. A., Ali, A., and Sundaraj, K. 2018. A literature review on NoSQL database for big data processing. International Journal of Engineering & Technology. 7, 2 (Jun. 2018), 902–906. DOI: <https://doi.org/10.14419/ijet.v7i2.12113>.

[5] Tarca, A. L., Carey, V. J., Chen, X., Romero, R., and Drăghici, S. 2007. Machine Learning and Its Applications to Biology. PLoS Comput. Biol. 3, 6 (June. 2007), 116. DOI: <https://doi.org/10.1371/journal.pcbi.0030116>.

[6] Chervonenkis, A. Y. 2013. Early History of Support Vector Machines. Empirical Inference, Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, 13–20. DOI: <https://doi.org/10.1007/978-3-642-41136-6_3>.

[7] Berkson, J. 1944. Application of the Logistic Function to Bio-Assay. Journal of the American Statistical Association. 39, 227 (Sep. 1944), 357–365. DOI: <https://doi.org/10.1080/01621459.1944.10500699>.

[8] Safavian, S.R. and Landgrebe, D. 1991. A survey of decision tree classifier methodology. IEEE Transactions on Systems, Man, and Cybernetics. 21, 3 (1991), 660–674. DOI: <https://doi.org/10.1109/21.97458>.

[9] Breiman, L. 2001. Random Forests. Machine Learning. 45, 1 (2001), 5–32. DOI: <https://doi.org/10.1023/A:1010933404324>.

[10] Leung, K.M. 2007. Naive bayesian classifier. Polytechnic University Department of Computer Science Finance and Risk Engineering. (2007).