


# A Machine Learning-Based Intelligent System for Predicting Diabetes

Nabila Shahnaz Khan, Department of Computer Science and Engineering, Military Institute of Science and Technology (MIST), Dhaka, Bangladesh

Mehedi Hasan Muaz, Department of Computer Science and Engineering, Military Institute of Science and Technology (MIST), Dhaka, Bangladesh

Anusha Kabir, Department of Computer Science and Engineering, Military Institute of Science and Technology (MIST), Dhaka, Bangladesh

Muhammad Nazrul Islam, Department of Computer Science and Engineering, Military Institute of Science and Technology (MIST), Dhaka, Bangladesh

 <https://orcid.org/0000-0002-7189-4879>

## ABSTRACT

In this era of technological growth, the diagnosis of diseases and finding cures, personal health parameter management and predicting the possibility of susceptibility to some diseases have become accessible and easy. Although all over the world millions of people are falling victim to diabetes, in most of the cases they are not even aware of their situation due to the silent nature of diabetes. Therefore, the objective of this research is to propose an intelligent system based on a machine learning algorithm to improve the accuracy of predicting diabetes. To attain this objective, an algorithm was proposed based on Naïve Bayes with prior clustering. Second, the performance of the proposed algorithm was evaluated using 532 data related to diabetic patients. Finally, the performance of the existing Naïve Bayes algorithm was compared with the proposed algorithm. The results of the comparative study showed that the improvement in the accuracy has been made apparent for the proposed algorithm.

## KEYWORDS

Algorithm, Clustering, Diabetes, Health Informatics, Machine Learning, Naïve Bayes Classifier

## 1. INTRODUCTION

Diabetes is a lifestyle disease with no cure. Its lifelong existence in the body gradually initiates other diseases and decays different organs. Most of the times diabetes remains hidden in the patient's body and do not show any syndrome. Currently,

DOI: 10.4018/IJBDAH.2019070101

This article, originally published under IGI Global's copyright on January 17, 2020 will proceed with publication as an Open Access article starting on January 20, 2021 in the gold Open Access journal, International Journal of Big Data and Analytics in Healthcare (converted to gold Open Access January 1, 2021), and will be distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

Diabetes Mellitus is spreading at an alarming rate. According to WHO (2016), about 422 million people were living with diabetes in 2014 and this number was estimated to increase to about 693 million by the year 2045 (Arnhold et al., 2014; Cho et al., 2018). About half of all people (49.7%) living with diabetes are undiagnosed, and the estimated healthcare cost of diabetic patients was USD 850 billion in 2017 (Cho et al., 2018). The WHO report also highlighted that 3.7 million deaths have been caused by diabetes (WHO, 2016). This alarming growth rate of diabetes is putting peoples' lives at risk worldwide, which is why it has become one of the foremost health concerns. Again, though many people have type 2 diabetes but still, its existence is not evident to them (Jourdan, 2012). Diabetes can be diagnosed through various types of blood tests which do not come handy and neither are they cheap. So, the rate of unawareness remains high. As diabetes is a hidden epidemic and a global health issue, predicting diabetes at an early stage or its probability beforehand gives the patients scopes to rebuild their lifestyle and food routine to save their lives. Thus, the development of an intelligent system for predicting the possibility of being diabetic becomes essential for the general people.

In the field of machine learning, the family of Naïve Bayes classifiers is regarded as one of the most common ways of predictive categorization using the probabilistic assumption of independent features (Wu et al., 2008). This group of algorithms has found their way into various fields like text categorization and analyzing documents (Chen et al., 2008; Schneider, 2005; Kibriya et al., 2004). In this research, Naïve Bayes classifier has been chosen primarily to propose an intelligent system for diabetic prediction despite the availability of other algorithms due to some specific reasons that includes: firstly, Naïve Bayes is remarkably simple to implement, has low computational complexity (Elkan et al., 1997) and provides very good accuracy (Ting et al., 2011). Secondly, Naïve Bayes classifier regards all the features equally for prediction. Finally, Naïve Bayes classifier is well known for its wide range application in healthcare prediction systems (Langerizadeh et al., 2016; Bhuvaneshwari et al., 2012). Considering all these conveniences, Naïve Bayes classifier was considered as more prominent for implementing the diabetes predicting system compared to other machine learning algorithms.

Therefore, the objective of this research is to propose a Naïve Bayes based intelligent system for predicting diabetes. It is also worth mentioning here that an earlier version of this research was published in (Khan et al., 2017), where a mobile application was built for predicting diabetes based on the existing Naïve Bayes algorithm. In this research, an advanced algorithm is proposed to predict the possibility of being diabetic or non-diabetic more accurately and efficiently.

Later sections of this paper are organized as follows. The related works are briefly introduced in Section II. The methodology which was followed throughout this research is presented in Section III. Later in Section IV, the developed algorithm is discussed along. Section V discussed the performance of the algorithm. The comparison of performance between existing Naïve Bayes classifier and the newly developed

algorithm are discussed in Section VI. The final section presents the main outcomes, practical implications, limitations of this work and concluding remarks.

## 2. LITERATURE REVIEW

Use of machine learning is prominent in the field of health informatics (Holzinger, 2014; Inan et al., 2018) and predicting the possibility of various diseases (Agarwal et al., 2018; Omar et al., 2019). This section briefly discusses the related work focusing to the techniques of predicting diabetes.

Some studies have been carried out for analyzing the risk factors of diabetes. Cafazzo et al. (2012) developed an application for (type 1) diabetic patients to monitor their blood glucose rate. A patient-centered framework for personalized health care has been proposed by Chawla & Davis (2013). In (Mougiakakou et al., 2010), a system SMARTDIAB consisting two units like patient unit and patient management unit was developed to provide management, treatment, and monitoring support of type 1 diabetes mellitus.

Some other studies have been carried out to predict the risk of diabetes based on Naïve Bayes. For example, Guo et al. (2012) tried to predict the possibility of developing type 2 diabetes applying Bayes Network using Weka software and the results were quite accurate. Similarly, Mani et al. (2012) strived to forecast type 2 diabetes risk using data from EMR based on the Gaussian Naïve Bayes and Logistic Regression algorithms. In another work, Parthiban et al. (2011) proposed a prediction model to predict the risk of developing heart disease of diabetic patients based on Naïve Bayes data mining classifier technique. Artificial intelligence-based techniques were also used to predict the diabetic. For example, in (Lee et al., 2011), a diabetic decision support system was developed based on the fuzzy expert system. In another study, El-Sappagh et al. (2018) proposed a case-based preparation framework for case-based reasoning (CBR) systems which converts the electronic health record medical data into fuzzy CBR knowledge. Barakat et al. (2010) proposed an intelligent representation of support vector machines (SVMs) that produces a comprehensive rule set matching the results of other related medical studies for the diagnosis of diabetes. Again, both the Generalized Discriminant Analysis (GDA) and Least Square Support Vector Machine (LS-SVM) techniques were used to diagnose diabetes in (Polat, 2008).

A few studies have been focused to explore the possible factors of diabetic prediction. For example, Breault et al. (2002) highlighted the factors like age, HbA1c, hypertension, and gender to analyze the data of diabetic patients and used the data mining technique. Patient habits and A1c data was used to diagnosing the possibility of being diabetic in (Sakshaug et al., 2014). Similarly, some common factors of type2 diabetes like gene factors, excessive weight gain, age, unhealthy food habit and lifestyle and polluted environment were pointed out by Marx (2002).

A few studies have been focused on comparing the performance of different diabetic prediction models/systems. For example, in (Meng et al., 2013), logistic regression, artificial neural network (ANN) and decision tree-based modes to predict diabetic

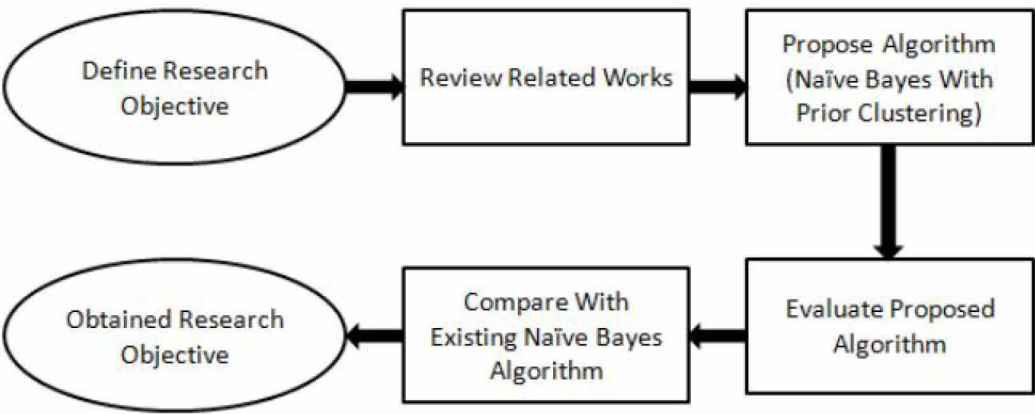
were compared. Another study shows comparison of different classifiers for diabetes prediction, that includes artificial neural networks, decision tree, logistic regression and Naïve Bayes (Nai-arun & Mounghmai, 2015), while Sarwar and Sharma (2014) analyzed the efficiency of algorithms implementing Naïve Bayes, artificial neural network (ANN), and k-nearest neighbors (KNN) to predict diabetics. Again, Temurtas et al. (2009) compare the multilayer neural network (MLNN) diabetes diagnosis with the Levenberg-Marquardt (LM) algorithm and probabilistic neural network (PNN) separately. As an outcome, their study showed an accuracy of 82.37% and 78.13% respectively while Pima Indian Diabetes data set was used.

In sum, the literature survey showed that there are several techniques for disease prediction and Naïve Bayes is a mentionable one among them. The accuracy of Naïve Bayes was found to be better than other existing algorithms like SMO, IB1, and Id3 (Cufoglu et al., 2009). Moreover, Naïve Bayes was quicker to converge than other classifier algorithms. Thus, this research work focuses on Naïve Bayes technique to enhance the accuracy of diabetes prediction.

### 3. RESEARCH METHODOLOGY

An overview of the research methodology has been presented in Figure 1. Firstly, the objective of this research was defined that is to develop a diabetes predictor algorithm with enhanced performance compared to the existing algorithms. After that, the related works were reviewed through a systematic literature review approach. Summary of the reviews has been presented in the previous section. Next, a dataset related to diabetes was selected. Thereafter, an algorithm was proposed based on Naïve Bayes Classifier with Prior Clustering to predict if the user is diabetic or not. Later on, the algorithm was evaluated using the selected dataset. Finally, the proposed algorithm was compared with the existing Naïve Bayes algorithm to show the effectiveness of the proposed algorithm for detecting diabetes using the selected dataset.

Figure 1. An overview of the research



## 4. PROPOSED ALGORITHM FOR THE INTELLIGENT PREDICTION SYSTEM

An algorithm was proposed to develop the intelligent system for predicting diabetes. The algorithm (Algorithm 1) was proposed primarily based on Naïve Bayes classifier. To improve the accuracy and efficiency of the algorithm, unsupervised classification has been integrated with the existing Naïve Bayes Classifier (supervised learning) as shown in Algorithm 1. This algorithm works in two stages. Firstly, the BSAS (Basic Sequential Algorithmic Scheme) approach was implemented to cluster similar datasets [Line 1-15]. Secondly, the Naïve Bayes algorithm was implemented on the specific cluster which contains the test data (user input) [Line 16-25]. A set of common features related to prediction of diabetes were selected for training and testing purpose; that includes: Plasma glucose concentration, diastolic blood pressure (mm Hg), triceps skin fold thickness (mm), body mass index (weight in kg/ height in m<sup>2</sup>), and age (Schrier, 2000; Ohlson et al., 1985; Hoda & Cheng, 2017).

### 4.1. Clustering Using BSAS

Clustering is a type of unsupervised learning where the model divides the given input data into different clusters based on the similarity of the features and it deals with finding a structure in a collection of unlabeled data. So, similar types of input data are expected to remain in the same cluster while data having distinguished features are contained in different clusters. A cluster is, therefore, a collection of objects which are “similar” to each other and are “dissimilar” to the objects belonging to other clusters (Karim, 2017). In Figure 2 we can see how similar types of data are classified into different clusters. Different types of sequential clustering are existent, such as BSAS (Basic Sequential Algorithmic Scheme), MBSAS (Modified Basic Sequential Algorithmic Scheme), K-means; but in this proposed algorithm BSAS has been used to cluster the dataset as it's the basic one.

In the conventional technique of Naïve Bayes Classifier, all the data present in the dataset are supposed to be used to train the machine and then the machine would give output for the current input. But with the improved version, the machine will first divide all the data present in the dataset into different clusters. In Algorithm 1, the dissimilarity between the data was calculated based on five features. The clustering algorithm was implemented on the dataset taking different values of  $q$  (maximum allowable clusters) and  $\theta$  (threshold of dissimilarity) and it was seen that the accuracy was maximum for  $q = 5$  and  $\theta = 17$ . So, in this algorithm, a maximum of five clusters are allowed and used threshold value of dissimilarity is 17. Then based on similarity, the machine will assign the user input data to any one of the clusters and the further work will be done using the data of that specific cluster. For instance, if used data in this algorithm could be divided into five clusters (cluster 0, 1, 2, 3, 4) as shown in Figure 2 and the test data belonged to cluster 3, then further operation would be done using the data belonging to cluster 3 only.

Figure 2. Dividing data into different clusters



In Algorithm 1, Line [1-15] represent the section where clustering has been done and the symbols used here represent the following meaning:

Dataset,  $X = \{x_1, x_2, \dots, x_n\}$

$d(x, C)$  = Dissimilarity between feature vector  $x$  and cluster  $C$

$\theta$  = Threshold of dissimilarity

$q$  = Maximum allowable clusters

$m$  = Current cluster no. after each step

$N$  = Total no. of training data

$X_{\text{input}}$  = Input feature

Clustering has been done using the following steps:

- Initially, the number of clusters is considered to be 1 [Line 3];
- The first input data belonging to the training dataset is assigned to the first cluster [Line 4];

**Algorithm 1. Improved algorithm for predicting the possibility of diabetes**

```

1. features  $\leftarrow \{\text{Glucose, BloodPressure, SkinThickness, BMI, Age}\}$ 
2. classes  $\leftarrow \{\text{Diabetic, Nondiabetic}\}$ 
3.  $m = 1$ 
4.  $C_m = \{x_1\}$ 
5. for  $i = 2$  to  $N$  do
6.   Find  $C_k : d(x_i, C_k) = \min_{1 \leq j \leq m} d(x_i, C_j)$ 
7.   if  $d(x_i, C_k) > \theta$  AND  $m < q$  then
8.      $m = m + 1$ 
9.      $C_m = \{x_i\}$ 
10.  else
11.     $C_k = C_k \cup \{x_i\}$ 
12.  end if
13. end for
14. Find  $C_k : d(x_{input}, C_k) = \min_{1 \leq j \leq m} d(x_{input}, C_j)$ 
15.  $C_{input} = C_k$ 
16. for each  $c_i$  in classes do
17.   Calculate  $\mu$  and  $\sigma$  for Glucose, BloodPressure, SkinThickness,
    BMI, Age for data contained in cluster  $C_{input}$ 
18. end for
19. for each case in test data of cluster  $C_{input}$  do
20.   for each  $c_i$  in classes do
21.      $posterior(c_i) \leftarrow \frac{P(c_i) \prod P(x_i | c_i)}{\text{evidence}}$  where  $x_i \in \text{features}$ 
22.   end for
23.  $max\_posterior \leftarrow \max(posterior(c_i \in \text{classes}))$ 
24.  $resulting\_class \leftarrow c_i$  such that  $c_i \in \text{classes}$  and  $posterior(c_i) = max\_posterior$ 
25. end for

```

- Then for the remaining training data  $\{x_2, x_3, \dots, x_n\}$ , dissimilarity between each data  $x_i$  where  $x_i \in \{x_2, x_3, \dots, x_n\}$  and all the existing clusters at that moment are calculated. If the calculated distance is greater than threshold value  $\theta$  and the current number of clusters  $m$  is less than the maximum possible number of clusters then a new cluster is created and  $m$  is incremented by 1. The data  $x_i$  is then assigned to the newly formed cluster [Line 5-13];

- After clustering all the given training data, the test data (user input) is assigned to the specific cluster having the minimum dissimilarity with the test data [Line 14-15].

## 4.2. Naïve Bayes on Specific Cluster

Naïve Bayes, a conditional probability model (Leung, n.d.), was applied to the dataset belonging to only that cluster containing the user input. Thus, the machine can be trained using only similar types of data and the possibility of misleading the machine will decrease. In Algorithm 1, Line [16-25] represents how the Naïve Bayes has been implemented. Given a problem instance to be classified, represented by a vector  $x = \{x_1, x_2, \dots, x_n\}$ , Naïve Bayes assigns probabilities  $p(C_k | x_1, \dots, x_n)$  for each of  $K$  possible outcomes or classes. Here,  $x = \{x_1, x_2, \dots, x_n\}$  representing some  $n$  features (independent variables). Using Bayes theorem, the conditional probability can be decomposed as:

$$p(C_k | x) = \frac{p(C_k) p(x | C_k)}{p(x)} \quad (1)$$

where:

$$p(x | C_k) = p(x_1 | C_k) * p(x_2 | C_k) * \dots * p(x_n | C_k)$$

The equation can be written as:

$$\text{posterior} = \frac{\text{prior} * \text{likelihood}}{\text{evidence}} \quad (2)$$

Again, in Naïve Bayes (Lowd & Domingos, 2005), if class is represented by  $C$  and features are represented using  $\{x_1, x_2, \dots, x_n\}$ , then the joint probability is:

$$P(C | x_1, x_2, \dots, x_n) = P(C) \prod_{i=1}^n P(x_i | C) \quad (3)$$

Here,  $P(C | x_1, x_2, \dots, x_n)$  represents the probability of being in class  $C$  while the features are  $\{x_1, x_2, \dots, x_n\}$ . In this research, Gaussian Naïve Bayes has been used as all the features are continuous. The second stage (Naïve Bayes on Specific Cluster) of the proposed algorithm [Algorithm 1] is executed using two phases: training phase and testing phase. In the training phase [line number 16-18 in Algorithm 1], initially the system taking the training data on as input to enhances its knowledge base; and then, using Gaussian Naïve Bayes the mean ( $\mu$ ) and variance ( $\sigma^2$ ) were calculated for all the features of each class. In the second phase, the system is tested [line 19-25

in Algorithm 2] using the following steps: Firstly, considering the test data as input, the conditional probability  $P(x|C)$  of each feature are calculated using Equation 4 as the data are in normal distribution:

$$P(x | C) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (4)$$

Next, posterior probability  $P(C|x)$  of belonging to each predefined class having features  $\{x_1, x_2, \dots, x_n\}$  is calculated using Equation 3. Finally, according to the MAP (maximum a posteriori) decision rule, the class having the maximum posterior probability is taken to be the resultant class for that data set.

## 5. EVALUATE THE ALGORITHM

This section unfolds the evaluation of the performance of the proposed intelligent system (algorithm) by introducing the selected dataset and simulating the proposed algorithm to assess its performance in terms of accuracy, sensitivity, and, specificity.

### 5.1. Selection of Dataset

The Pima Indian Diabetes Dataset (Pima, n.d.) was selected to evaluate the effectiveness and efficiency of the proposed algorithm because all the five features required to evaluate the algorithm was available in this dataset. It is a freely available open source online dataset, as well as used in different research works (Frank & Hall, 2003; Tong & Koller, 2000). This dataset contains a total of 768 data. All participants in this dataset are females of Pima Indian heritage and are at least 21 years old. After processing the dataset, 532 data were found to be reliable to use for further implementation.

### 5.2. Simulation of the Algorithm

An example is discussed here to show how the proposed algorithm works to anticipate a probable condition of the patient of being diabetic or nondiabetic and classify accordingly. In this example, a random case was considered as the test data while the rest of 531 data consisted of the training dataset. After running the clustering portion of the algorithm, the total 532 data were divided into five clusters where the first cluster contained 77 data, 2nd one contained 162 data, 3rd one contained 87 data, 4th one contained 149 data and the 5th cluster consisted of 57 data. It was seen that the test case was contained in the fourth cluster (shown as cluster 3 in Figure 2). So, only the 148 data (excluding the test case) belonging to this cluster were considered for training the machine. Now, in the training phase, it was supposed that for a particular feature from the training set, mean is  $\mu_1$  and variance is  $\sigma_1$  in case of diabetic and  $\mu_2$  and  $\sigma_2$  for class nondiabetic. The mean and variances are shown in Table 1.

Here in testing phase, the values of the features of a test case are age = '23' years, BMI = '30.4' kg/m<sup>2</sup>, Triceps skin fold thickness = '31', Plasma glucose concentration

**Table 1. Mean and variance calculation for continuous features**

Feature	$\mu_1$	$\sigma_1$	$\mu_2$	$\sigma_2$
Glucose	88.375	7795.5	90.493	8234.329
Blood Pressure	62.0	4009.0	64.536	4228.786
Skin Thickness	26.25	778.25	22.5	547.986
BMI	33.425	1167.688	28.724	829.759
Age	30.875	996.0	26.186	702.357

= ‘93’, Diastolic blood pressure = ‘70’ mmHg. Now, for class diabetic, putting these values in Equation 4 we get,  $P(\text{age}|\text{diabetic}) = 0.0123$ ,  $P(\text{BMI}|\text{diabetic}) = 0.0116$ ,  $P(\text{skinThickness}|\text{diabetic}) = 0.0141$ ,  $P(\text{glucose}|\text{diabetic}) = 0.0045$ ,  $P(\text{pressure}|\text{diabetic}) = 0.0063$ . Similarly, for nondiabetic,  $P(\text{age}|\text{nondiabetic}) = 0.0149$ ,  $P(\text{BMI}|\text{nondiabetic}) = 0.0138$ ,  $P(\text{skinThickness}|\text{nondiabetic}) = 0.0160$ ,  $P(\text{glucose}|\text{nondiabetic}) = 0.0044$ ,  $P(\text{pressure}|\text{nondiabetic}) = 0.0061$ . The prior probabilities are  $P(\text{diabetic})=0.0540$  and  $P(\text{nondiabetic})=0.9459$ . Now using Equation 3 yields the following posterior probabilities which will help to predict the class in which the test case belongs. After calculation,  $P(\text{diabetic}|x) = 0.0353$  and  $P(\text{nondiabetic}|x) = 0.9647$ . As the posterior probability of being nondiabetic is high, the case is assumed to be under class nondiabetic.

### 5.3. Performance Evaluation

Performance of the proposed algorithm was evaluated and the probability of error of the classifier was calculated to decide the efficiency of the proposed algorithm. With a view to estimating the accuracy of the proposed algorithm and evaluating it in terms of sensitivity and specificity, the algorithm was applied on 532 processed data of ‘PIMA Indian Diabetes Dataset’. For evaluating the system, four performance measures were considered, which are True Positive (the number of cases correctly identified as patient), True Negative (the number of cases correctly identified as healthy), False Positive (the number of cases incorrectly identified as patient) and False Negative (the number of cases incorrectly identified as healthy). Table 2 represents the confusion matrix showing the values of four performed measures.

**Table 2. Confusion Matrix**

	Actual Diabetic	Actual Nondiabetic	Total
Test Diabetic	True Positive (TP) = 95	False Positive (FP) = 39	134
Test Nondiabetic	False Negative (FN) = 82	True Negative (TN) = 316	398
Total	177	355	532

The accuracy, specificity, and sensitivity of the proposed algorithm was calculated based on the values of Table 2 and using the Equations 5, 6, and 7, respectively. Accuracy is the measure of how correctly the algorithm differentiates between healthy human and diabetic patients; Sensitivity is the measure of how correctly the diabetic patients are identified while specificity represents how correctly healthy humans are identified. The obtained values are Accuracy = 77.26%; Sensitivity = 53.67%; Specificity = 89.01% as shown in Figure 3:

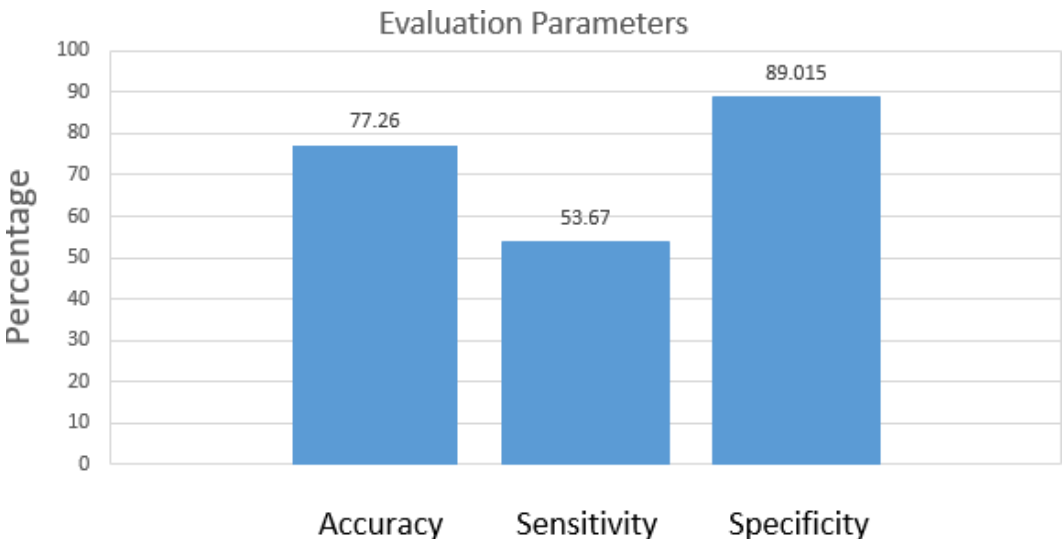
$$\text{Accuracy} = \frac{(TN + TP)}{(TN + TP + FN + FP)} \quad (5)$$

$$\text{Sensitivity} = \frac{TP}{(TP + FN)} \quad (6)$$

$$\text{Specificity} = \frac{TN}{(TN + FP)} \quad (7)$$

To estimate the possibility of error of the system leave-one-out-technique (Elisseff & Pontil, 2003) was used. In this technique, each time one observation from the whole dataset is taken as test data and the rest of the dataset is considered to be the training data (Online Resource, n.d.). Thus firstly, the first sample was used as test data and the rest  $n-1$  samples were used as training data. Then again, the second sample was used as test data while others were used as training data. Thus, the process was repeated

Figure 3. Result obtained by evaluating the proposed



$n-1$  times. For each case, while estimating the error of the test sample the fractional counting was used which uses the estimated probability of class membership of sample.  $\hat{P}(C_i|x)$  is the largest of probabilities for  $C_i$  class and sample  $x$ . In such a case, its probability of being erroneous is found by Equation 8:

$$\hat{P}(\varepsilon) = 1 - \hat{P}(C_i | x) \quad (8)$$

$$\hat{P}(\varepsilon) = \sum_{i=1}^n \hat{P}_i(\varepsilon | x) \quad (9)$$

Table 3 shows error calculation for two sample cases where class Diabetic is represented by 'A' and Nondiabetic is represented by 'B'. The total estimated error for  $n$  samples has been calculated using Equation 9 and the resultant estimated error was found to be  $\hat{P}(\varepsilon) = 0.204$ .

## 6. COMPARISON WITH EXISTING NAÏVE BAYES ALGORITHM

The proposed algorithm used the concept of the basic Naïve Bayes technique and incorporated the concept of prior clustering. This was chosen to exaggerate the level of accuracy and enhance the effectiveness of the existing algorithm. Thus, a comparative study was conducted to portray the effectiveness of the proposed algorithm in contrast to the existing Naïve Bayes Algorithm. The basic Naïve Bayes algorithm is presented in Algorithm 2.

Both the proposed and the existing algorithm was implemented using the same IDE and was executed on the 532-usable data of Pima Indian Dataset separately considering the similar set of features for both the cases. Again, for evaluating the algorithms, accuracy, sensitivity, specificity, and the probability of error of classifiers for the two algorithms were calculated following the same procedure (as discussed in Section 5.3). The results are shown in Table 4 and Figure 4.

The result showed that the *accuracy* of the proposed algorithm was increased by 10.34% which indicates that the percentage of correct assumptions by the system increases by an amount of 10.34%. The result related to *sensitivity* shows a rise in *sensitivity* by 53.11%, which indicates a significant increase in the ability to detect actual diabetes cases correctly. Although there has been a fall of *specificity* by 10.99%

**Table 3. Error estimation**

Test Case	True Class	$P(A   x)$	$P(B   x)$	$\hat{P}(\varepsilon   x)$
1	A	0.69	0.31	0.31
2	B	0.63	0.37	0.37

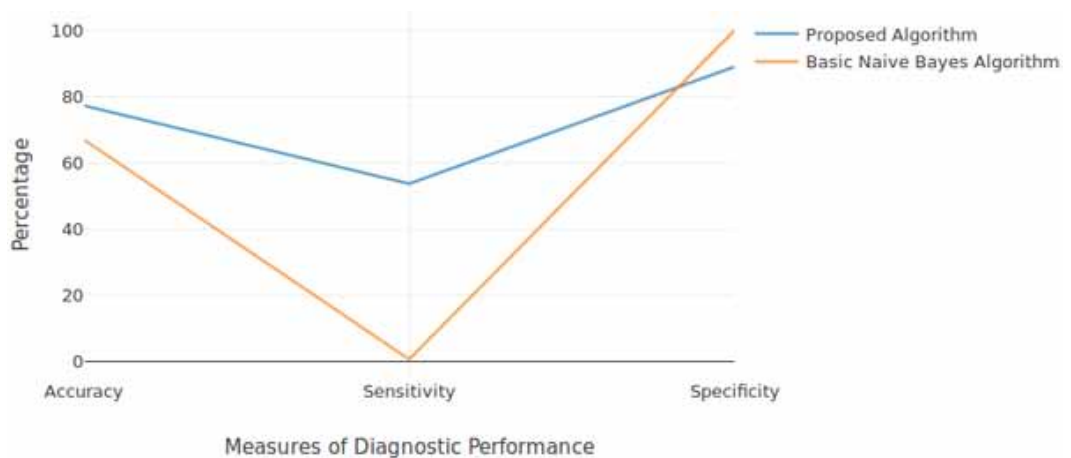
**Algorithm 2. Predicting the possibility of diabetes using basic Naïve Bayes**

1.	features $\leftarrow \{\text{Glucose, BloodPressure, SkinThickness, BMI, Age}\}$
2.	classes $\leftarrow \{\text{Diabetic, Nondiabetic}\}$
3.	for each $c_i$ in classes do
4.	Calculate $\mu$ and $\sigma$ for Glucose, BloodPressure, SkinThickness, BMI, Age
5.	end for
6.	for each case in test data do
7.	for each $c_i$ in classes do
8.	$posterior(c_i) \leftarrow \frac{P(c_i) \prod P(x_i   c_i)}{\text{evidence}}$ where $x_i \in \text{features}$
9.	end for
10.	$max\_posterior \leftarrow \max(posterior(c_i \in \text{classes}))$
11.	$resulting\_class \leftarrow c_i$ such that $c_i \in \text{classes}$ and $posterior(c_i) = max\_posterior$
12.	end for

**Table 4. Comparing the performance between two algorithms**

Algorithm	Accuracy	Sensitivity	Specificity	Probability of Error
Existing Algorithm (Naïve Bayes)	66.92%	0.56%	100%	0.185
Proposed Algorithm (Naïve Bayes with Prior Clustering)	77.26%	53.67%	89.015	0.204

**Figure 4. Difference in accuracy of the two algorithms**



of the proposed algorithm with respect to the existing Naïve Bayes algorithm which is quite evident as sensitivity and specificity are inversely related. The value of *specificity* was less significant compared to the overall improvement of performance. The probability of error of the proposed algorithm was increased by a small fraction of 0.019 than that of the existing algorithm. The comparison clearly indicates that the proposed algorithm gives better performance for diabetes prediction than the existing Naïve Bayes algorithm.

## 7. DISCUSSION AND CONCLUSION

To create awareness among people, different machine learning algorithms including the Naïve Bayes classifier have been in use for prediction of probabilities of diseases (including diabetes) since long. On assessment, as basic Naïve Bayes was found to provide an inadequate level of accuracy, an attempt to gain an expected level of accuracy and lesser probability of error was made. Therefore, in this research an intelligent prediction system is proposed to provide a better prediction of diabetes specifically in an easily accessible method and thereby contribute to create awareness among people about the disease. The Basic Sequential Clustering algorithm and Naïve Bayes technique were incorporated to develop the intelligent system. The proposed algorithm was then evaluated with Pima dataset and later compared to the existing algorithm. The comparative study showed that the proposed algorithm is better in terms of *accuracy* than that of the existing Naïve Bayes algorithm by 10.34%. There has been a slight fall in *specificity* by 10.99% and the probability of error of classifier by 0.019% but a dynamic rise in *sensitivity* by 53.11%. The significant increase in *accuracy* and *sensitivity* indicates that the proposed algorithm could be a better choice.

The study has some limitations as well. For example, insufficient amount of data of female patients was used in this research. Again, the algorithm was proposed considering only five features. However, the algorithm can be made to function with datasets containing other features which lie in the root of diabetes keeping necessary conditions unchanged. In such a case, the performance of the algorithm is subjected to evaluation.

The focus for future expansion would be the enhancement of the knowledge base by collecting a sufficiently large amount of data both from male and female patients of all age groups to improve the accuracy of the system. Again, comparing the proposed algorithm with the other existing machine learning algorithms to find the most appropriate one for predicting diabetes would be another option for future research.

In this paper, an algorithm has been proposed that can predict the state or possibilities of diabetes of a person more accurately than that of existing Naïve Bayes technique. The proposed system will assist any individual to know his/her probability of being diabetic even when at home without consulting a doctor. This reduces the efforts required to meet a physician in person alongside raising awareness.

As diabetes can remain unrevealed for a long time and the causes of diabetes are yet not certain, its explosion rate is growing day by day anomalously. In this hazardous situation, it is sincerely hoped that such a diabetes prediction system would be able to mitigate the explosion rate of diabetes by creating awareness among general people regarding diabetes.

## REFERENCES

- Agrawal, A., Agrawal, H., Mittal, S., & Sharma, M. (2018). Disease Prediction Using Machine Learning. In *Proceedings of the 3rd International Conference on Internet of Things and Connected Technologies (ICIOTCT)* (pp. 419–422). Academic Press.
- Arnhold, M., Quade, M., & Kirch, W. (2014). Mobile applications for diabetics: A systematic review and expert-based usability evaluation considering the special requirements of diabetes patients age 50 years or older. *Journal of Medical Internet Research*, 16(4), e104. doi:10.2196/jmir.2968 PMID:24718852
- Barakat, N., Bradley, A. P., & Barakat, M. N. H. (2010). Intelligible support vector machines for diagnosis of diabetes mellitus. *IEEE Transactions on Information Technology in Biomedicine*, 14(4), 1114–1120. doi:10.1109/TITB.2009.2039485 PMID:20071261
- Bhuvaneswari, R., & Kalaiselvi, K. (2012). Naïve Bayesian classification approach in healthcare applications. *International Journal of Computer Science and Telecommunications*, 3(1), 106–112.
- Breault, J. L., Goodall, C. R., & Fos, P. J. (2002). Data mining a diabetic data warehouse. *Artificial Intelligence in Medicine*, 26(1–2), 37–54. doi:10.1016/S0933-3657(02)00051-9 PMID:12234716
- Cafazzo, J. A., Casselman, M., Hamming, N., Katzman, D. K., & Palmert, M. R. (2012). Design of an mHealth app for the self-management of adolescent type 1 diabetes: A pilot study. *Journal of Medical Internet Research*, 14(3), e70. doi:10.2196/jmir.2058 PMID:22564332
- Chawla, N. V., & Davis, D. A. (2013). Bringing big data to personalized healthcare: A patient-centered framework. *Journal of General Internal Medicine*, 28(3), 660–665. doi:10.1007/s11606-013-2455-8 PMID:23797912
- Chen, J., Huang, H., Tian, S., & Qu, Y. (2009). Feature selection for text classification with Naïve Bayes. *Expert Systems with Applications*, 36(3), 5432–5435. doi:10.1016/j.eswa.2008.06.054
- Cho, N. H., Shaw, J. E., Karuranga, S., Huang, Y., da Rocha Fernandes, J. D., Ohlrogge, A. W., & Malanda, B. (2018). IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. *Diabetes Research and Clinical Practice*, 138, 271–281. doi:10.1016/j.diabres.2018.02.023 PMID:29496507
- Cufoglu, A., Lohi, M., & Madani, K. (2009). A comparative study of selected classifiers with classification accuracy in user profiling. In *Proceedings of the 2009 WRI World Congress on Computer Science and Information Engineering* (Vol. 3, pp. 708–712). IEEE. doi:10.1109/CSIE.2009.954

- El-Sappagh, S., Elmogy, M., Ali, F., & Kwak, K.-S. (2018). A case-base fuzzification process: Diabetes diagnosis case study. *Soft Computing*.
- Elisseeff, A., & Pontil, M. (2003). Leave-one-out error and stability of learning algorithms with applications. *NATO Science Series Sub Series Iii Computer and Systems Sciences*, 190, 111–130.
- Elkan, C. (1997). *Naïve bayesian learning*. University of California.
- Frank, E., & Hall, M. (2003). Visualizing class probability estimators. In *Proceedings of the European Conference on Principles of Data Mining and Knowledge Discovery* (pp. 168–179). Springer.
- Guo, Y., Bai, G., & Hu, Y. (2012). Using bayes network for prediction of type-2 diabetes. In *Proceedings of the 2012 International Conference for Internet Technology and Secured Transactions* (pp. 471–472). IEEE.
- Hoda, S. A., & Cheng, E. (2017). *Robbins basic pathology*. Oxford University Press.
- Holzinger, A. (2016). Interactive machine learning for health informatics: When do we need the human-in-the-loop? *Brain Informatics*, 3(2), 119–131. doi:10.1007/s40708-016-0042-6 PMID:27747607
- Inan, T. T., Samia, M. B. R., Tulin, I. T., & Islam, M. N. (2018). A Decision Support Model to Predict ICU Readmission through Data Mining Approach. In *Proceedings of the 22nd Pacific Asia Conference on Information Systems (PACIS 2018)*. Academic Press.
- Jourdan, T. (2012). Hidden diabetes. Netdoctor. Retrieved from <http://www.netdoctor.co.uk/conditions/diabetes/news/a10473/hidden-diabetes/>
- Karim, M. R. (2017). *Scala and Spark for Big Data Analytics*. Packt Publishing.
- Khan, N. S., Muaz, M. H., Kabir, A., & Islam, M. N. (2017). Diabetes Predicting mHealth Application Using Machine Learning. In *Proceedings of the 2017 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE)* (pp. 237-240). IEEE. doi:10.1109/WIECON-ECE.2017.8468885
- Kibriya, A. M., Frank, E., Pfahringer, B., & Holmes, G. (2004). Multinomial Naïve bayes for text categorization revisited. In *Proceedings of the Australasian Joint Conference on Artificial Intelligence* (pp. 488–499). Springer. doi:10.1007/978-3-540-30549-1\_43
- Langarizadeh, M., & Moghbeli, F. (2016). Applying Naïve Bayesian networks to disease prediction: A systematic review. *Acta Informatica Medica*, 24(5), 364. doi:10.5455/aim.2016.24.364-369 PMID:28077895

- Lee, C.-S., & Wang, M.-H. (2011). A fuzzy expert system for diabetes decision support application. *IEEE Transactions on Systems, Man, and Cybernetics. Part B, Cybernetics*, 41(1), 139–153. doi:10.1109/TSMCB.2010.2048899 PMID:20501347
- Leung, K. M. (n.d.). Naïve bayesian classifier. Polytechnic University Department of Computer Science/Finance and Risk Engineering. *Lecture Notes*.
- Lowd, D., & Domingos, P. (2005). Naïve Bayes models for probability estimation. In *Proceedings of the 22nd international conference on Machine learning* (pp. 529–536). ACM.
- Mani, S., Chen, Y., Elasy, T., Clayton, W., & Denny, J. (2012). Type 2 diabetes risk forecasting from EMR data using machine learning. *AMIA ... Annual Symposium Proceedings - AMIA Symposium*. AMIA Symposium, 606. PMID:23304333
- Marx, J. (2002). *Unraveling the causes of diabetes*. American Association for the Advancement of Science. doi:10.1126/science.296.5568.686
- Meng, X.-H., Huang, Y.-X., Rao, D.-P., Zhang, Q., & Liu, Q. (2013). Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. *The Kaohsiung Journal of Medical Sciences*, 29(2), 93–99. doi:10.1016/j.kjms.2012.08.016 PMID:23347811
- Mougiakakou, S. G., Bartsocas, C. S., Bozas, E., Chaniotakis, N., Iliopoulou, D., Kouris, I., & Tsoukalis, A. (2010). SMARTDIAB: A communication and information technology approach for the intelligent monitoring, management and follow-up of type 1 diabetes patients. *IEEE Transactions on Information Technology in Biomedicine*, 14(3), 622–633. doi:10.1109/TITB.2009.2039711 PMID:20123578
- Nai-arun, N., & Moungrmai, R. (2015). Comparison of classifiers for the risk of diabetes prediction. *Procedia Computer Science*, 69, 132–142. doi:10.1016/j.procs.2015.10.014
- Ohlson, L.-O., Larsson, B., Svärdsudd, K., Welin, L., Eriksson, H., Wilhelmsen, L., & Tibblin, G. (1985). The influence of body fat distribution on the incidence of diabetes mellitus: 13.5 years of follow-up of the participants in the study of men born in 1913. *Diabetes*, 34(10), 1055–1058. doi:10.2337/diab.34.10.1055 PMID:4043554
- Omar, K. S., Mondal, P., Khan, N. S., Rizvi, M. R. K., & Islam, M. N. (2019). A Machine Learning Approach to Predict Autism Spectrum Disorder. In *Proceedings of the 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)* (pp. 1-6). IEEE. doi:10.1109/ECACE.2019.8679454
- Online Resource. (n.d.). Machine Learning corner. Retrieved from <https://mlcorner.wordpress.com/tag/leave-one-out/>
- Parthiban, G., Rajesh, A., & Srivatsa, S. K. (2011). Diagnosis of heart disease for diabetic patients using Naïve Bayes method. *International Journal of Computers and Applications*, 24(3), 7–11. doi:10.5120/2933-3887

- Pattekari, S. A., & Parveen, A. (2012). Prediction system for heart disease using Naïve Bayes. *International Journal of Advanced Computer and Mathematical Sciences*, 3(3), 290–294.
- Pima. (n.d.). Pima Indians Diabetes Database. Retrieved from <https://kaggle.com/uciml/pima-indians-diabetes-database>
- Polat, K., Güneş, S., & Arslan, A. (2008). A cascade learning system for classification of diabetes disease: Generalized discriminant analysis and least square support vector machine. *Expert Systems with Applications*, 34(1), 482–487. doi:10.1016/j.eswa.2006.09.012
- Sakshaug, J. W., Weir, D. R., & Nicholas, L. H. (2014). Identifying diabetics in Medicare claims and survey data: Implications for health services research. *BMC Health Services Research*, 14(1), 150. doi:10.1186/1472-6963-14-150 PMID:24693862
- Sarwar, A., & Sharma, V. (2014). Comparative analysis of machine learning techniques in prognosis of type II diabetes. *AI & Society*, 29(1), 123–129. doi:10.1007/s00146-013-0456-0
- Schneider, K.-M. (2005). Techniques for improving the performance of Naïve bayes for text classification. In *Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 682–693). Springer. doi:10.1007/978-3-540-30586-6\_76
- Schrier, R. W. (2000). Effect of blood pressure control on diabetic microvascular complications in patients with hypertension and type 2 diabetes. *Diabetes Care*, 23, B54–B64. PMID:10860192
- Temurtas, H., Yumusak, N., & Temurtas, F. (2009). A comparative study on diabetes disease diagnosis using neural networks. *Expert Systems with Applications*, 36(4), 8610–8615. doi:10.1016/j.eswa.2008.10.032
- Ting, S. L., Ip, W. H., & Tsang, A. H. (2011). Is Naïve Bayes a good classifier for document classification. *International Journal of Software Engineering and Its Applications*, 5(3), 37–46.
- Tong, S., & Koller, D. (2000). Restricted bayes optimal classifiers. In *AAAI/IAAI* (pp. 658–664). AAAI Press.
- WHO. (2016). Global report on diabetes: World Health Organization. Retrieved from [https://scholar.google.co.in/scholar?hl=en&as\\_sdt=0%2C5&q=Global+report+on+diabetes%3A+World+Health+Organization&btnG=](https://scholar.google.co.in/scholar?hl=en&as_sdt=0%2C5&q=Global+report+on+diabetes%3A+World+Health+Organization&btnG=)
- Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., & Philip, S. Y. et al. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1), 1–37. doi:10.1007/s10115-007-0114-2

*Nabila Shahnaz Khan has graduated from the CSE Department of Military Institute of Science and Technology (MIST) in 2018 and is currently working as a lecturer in MIST.*

*Mehedi Hasan Muaz is currently working as a software engineer and pursuing MSc Engineering in Computer Science and Engineering at the Military Institute of Science and Technology.*

*Anusha Kabir has graduated from the CSE Department of Military Institute of Science and Technology and currently working at a telecommunication company in Bangladesh.*

*Muhammad Nazrul Islam (PhD) is an Associate Professor in the Department of Computer Science and Engineering at the Military Institute of Science and Technology (MIST), Mirpur Cantonment, Dhaka, Bangladesh. He was awarded a Ph.D. in Information Systems from Åbo Akademi University (Finland) in 2014 and a M.Sc. in Computer Engineering from Politecnico di Milano (Italy) in 2007. Before joining MIST, he worked as a visiting teaching fellow at Uppsala University, Sweden, and as a postdoctoral research fellow at Åbo Akademi University, Finland. His research interests are focused on human-computer interaction, health informatics, military information systems, information systems usability, user experience, and computer semiotics. He is the author of about 80 peer-reviewed publications in international journals and conferences. He is a member of the IEEE and the IEB (Engineering Institute of Bangladesh).*