

Diabetes diagnosis system using modified Naive Bayes classifier

Jwan Kanaan Alwan¹, Dhulfiqar Saad Jaafar², Itimad Raheem Ali¹

¹Biomedical Informatics College, University of Information Technology and Communications, Baghdad, Iraq

²Rusafa 2 directorate, Ministry of Education, Baghdad, Iraq

Article Info

Article history:

Received Jan 15, 2022

Revised Aug 20, 2022

Accepted Sep 23, 2022

Keywords:

Data Mining

Diabetes diagnosis system

K-nearest neighbor algorithm

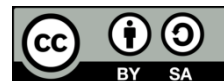
Modified Naive Bayesian

Traditional Naive Bayesian

ABSTRACT

In today's world, Diabetes is one of these diseases and is now a big growing health problem. The techniques of data mining have been widely applied to extract knowledge from medical databases. In this work, a Medical Diagnosis system of Diabetes is proposed for the diagnosis of diabetes in a manner that is rapid and cost-effective. three stages are involved in the proposed diabetes diagnosis system (DDS) including: dataset constructing, preprocessing and classification algorithm using traditional Naive Bayesian (TNB) and modified Naive Bayesian (MNB)). MNB Classifier is a modified NB that is used to enhance the accuracy of diagnosis, by adding a proposed modest model to help separate the overlapping diagnosis classes. The outcome showed that the accuracy of MNB classifier is generally higher than that of TNB classifier for all sets of features. An accuracy of about (63%) was achieved for the TNB model, whereas that of the MNB model is (100%). The experimental results showed that the MNB is better than the traditional NB in both two cases of constructed medical datasets; the first case of filling the missing values by experiences and the second case of filling missing values by K-nearest neighbor (KNN) algorithm.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Itimad Raheem Ali

Biomedical Informatics College, University of Information Technology and Communications

Baghdad, Iraq

Email: itimadra@uoitc.edu.iq

1. INTRODUCTION

Diabetes is an opportune disease which has a large wealth of data available and has huge complications. There is a need for a better and more accurate approach to diagnose the patients suffering from that disease [1]. Data Mining is one of the most vital and motivating areas of research with the objective of finding meaningful information from huge datasets. In the present era, data mining is becoming popular in the healthcare field because there is a need for an efficient analytical methodology for detecting unknown and valuable information in health data. However, it is still important to have approaches that can facilitate accurate diagnosis of the diseases. To this end, data mining is regarded as the most appropriate approach, because it is a process through which relevant and concealed information can be extracted from larger databases. This process can be executed using a wide range of available data mining techniques such as classification, association mining, predictive analysis, and clustering [2]-[5] an intelligent decision support system IDSS was designed to help patients and health-care providers [6]. The system was specially tailored to provide 24- hour adequate health-care services to diabetes type2 patients remotely, and provide decision support service to the health-care providers [7], [8]. The system was designed in a manner that allows the periodic updating of patients 'records. [9] addressed the adverse effect of missing value imputation in their work, and they also provided a solution for improvement during the evaluation of the performance of K-nearest neighbor (KNN) algorithm for the classification of Diabetes data. Their results showed that their

novel class-wise K-Nearest Neighbor (CKNN) algorithm significantly improved the accuracy of the CKNN method for the classification of diabetes dataset [10]. Their proposed approach is a two-level approach to data classification [11]. The initial phase of their method involves the extraction of the best feature set, then a new dataset is created as the best training dataset, and subsequently, the optimal feature set undergoes classification. Here, a classification mechanism is used for the classification of the testing data features with the new dataset; the Bayesian classifier is used for this purpose [12]-[14].

2. GENERAL DESCRIPTION OF THE PROPOSAL

This section provides a general description for proposes model. The model includes constructing the dataset, preprocessing, and classification alorithm used. Figure 1 shows the flowchart of overall diagnosis to be carried out using the proposed diabetes mellitus (DM)-diabetes diagnosis system (DDS). Each part of the flowchart is explained in the following subsections.

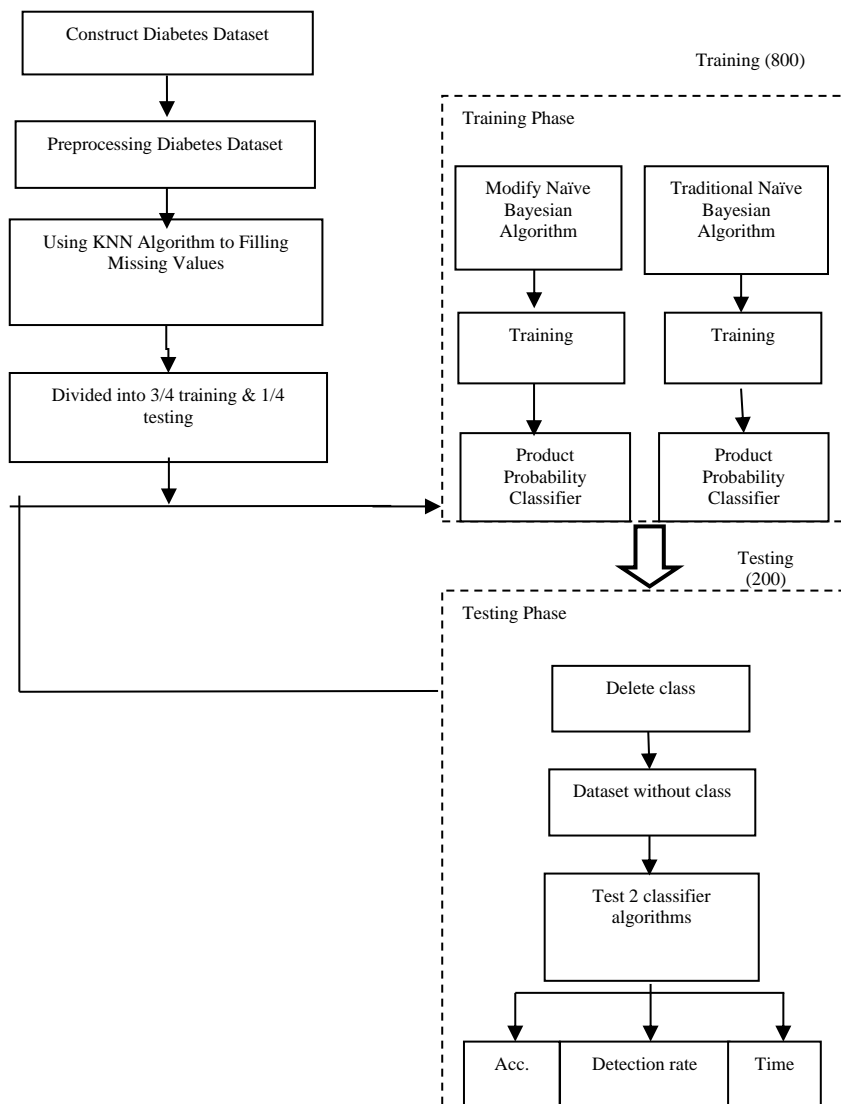


Figure 1. Weibull distribution of all filler concentrations

3. DESIGN AND IMPLEMENT THE PROPOSAL

In implementing the algorithms of data mining, the KNN, traditional Naive Bayes (TNB) and the modify Naive Bayes (MNB) [15]. Dataset divided into training and testing phase; several stages were included to obtain the end result of the process of implementation of the algorithms. We used accuracy ant training time for evaluation.

3.1. Construct diabetes dataset

The proposed system was applied on a constructed dataset of 1000 individuals from Baghdad society, covering the three classes. The data for this study were collected from the laboratory of Medical City Hospital and (the Specializes Center for Endocrinology and Diabetes-Al-Kindi Teaching Hospital) located in Iraq the detail information for data explained in Table 1. Before applying mining techniques, there are some preprocessing steps that were taken into account so as to prepare the data for training and testing. KNN is the critical preprocessing step aimed at filling the missing values, instead of guessing them by experience. Two algorithms of data mining were applied in this proposal TNB, and MNB classifier [16], [17]. The main tasks performed in the data preprocessing stage include: Removal of Redundancy, Noisy Data. Also, the preprocessing stage involved Feature Selection whereby only eight of the attributes collected (Age, Gender, HBA1C, TG, Urea, Chol, HDL and BMI) were considered, whereas, the less important features were ignored as their information gain is of no significance except for the Blood Sugar Level which was ruled out because it is the decisive factor in diabetes diagnosis [18], [19]. The data were applied using two different algorithms, and it was found that the level of accuracy was low. To this end, the following section explains how missing values were treated using a proposed KNN to increase the accuracy of diagnosis. In the proposed method, missing data were estimated and replaced using k-nearest neighbor algorithms. It is important the entire dataset be searched when the most similar instances are searched for by the KNN algorithms [20]-[22] based on (1).

$$E_v - E = \frac{h}{2.m} (k_x^2 + k_y^2) \quad (1)$$

All symbols that have been used in the equations should be defined in the following text.

Table 1. Classification Results of TNB and MNB Classifiers (testing of 200 patients)

Value	Classifier	TP	TN	FP	FN	Unknown1 (Pr-p)	Unknown2 (Pr-n)	Accuracy
Without Missing	TNB	123	3	25	38	0	11	63%
	MNB	150	39	0	0	11	0	100%
With Missing	TNB	119	3	31	38	2	9	62%
	MNB	120	20	30	19	3	8	71.5%

3.2. K-nearest neighbor method

With this simple algorithm, all available cases are stored and new cases are classified based on a similarity measure like distance functions [23]. The KNN was used to determine the missing values and to avoid the bad impact of the arbitrary speculations of the values. The imputations in the proposal begin with the distance similarity measure as in (2).

$$D = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2} \quad (2)$$

3.3. Naïve Bayesian classifiers

In Naïve Bayesian (NB) classifier [15] a set of probabilities (a priori, conditional, and posteriori) was found instead of constructing a set of classification rules. Assuming the features' conditional independence, as shown in (5) is used to compute the "conditional probabilities" $P(C_i|X)$ of the testing record at each class. Finally, in (3) is used to compute the "postpriori probability" $P(h|X)$ of the testing record at each class, with the class with the maximum posteriori probability h_{MAP} being labelled for the testing record based on (4).

$$P(h|X) = \frac{P(X|h)P(h)}{P(X)} \quad (3)$$

$$h_{MAP} = \arg \max_{h \in H} P(X|h)P(h) \quad (4)$$

Where, the set of hypotheses is denoted by H This is a statistically optimal classification rule which is commonly used to compare the performance of other classification algorithms [24]. Nevertheless, it is accompanied by some practical limitations like the need for prior knowledge about many probabilities, and a quiet high cost of computation [3].

$$P(C_j|X) > P(C_i|X) \quad (3)$$

FOR $1 < J < I$ AND $J \neq I$. USING EQ.

GETTING

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}, P(X|C_i) = \prod_{j=1}^N P(A_j|C_i) \tag{5}$$

3.3.1. The TNB Algorithm

In the traditional NB algorithm, each record in the training DDS will be classified through the following computations. The conditional probability P (ai Ci) for record's values a_j of its jth feature, that exists in the used set of features see Table 2, which has the value aj as the jth feature in class Ci. Assuming conditional independence of features, the "conditional probabilities" of the training record at each class are computed using (5). To avoid obtaining a zero result with this equation, if the frequency of (aj)as the jth feature of the testing record in class Ci equals to zero, then "increase the frequency of (aj) by one", also increase the frequency of class Ci in TrainingDDS by "a value that equals a number of available values of the jth feature". Finally, the "posteriori probability" of the training record at each class is computed using (2), and the class with the maximum posterior probability will be labelled for the training record according to (3).

Ambica *et al.* [11] implemented on TrainingDDS with each one of the sets (diabetic, non-diabetic, and predicted-diabetic) separately, and the classification of TestinDDS records has been simultaneously done with the implementation as shown in Table 3. Bayesian classifier is a simple classifier that requires just a one-time scanning of the data, which in turn increases the level of accuracy and speed for large DBs [25].

Table 2. The dataset of the DDS

ID	NO_Patient	GEN	AGE	Sugar	UREA	HB	CHOL	TG	HDL	BMI	CLASS
1	34325	M	58	6.7	20.8	9.1	6.6	2.9	1.1	33	Y
2	44835	M	60	10.8	2.1	7.6	3.3	1.7	0.9	36.6	Y
3	23972	F	56	7	4	9.2	4.1	0.6	1.3	30	Y
4	34301	F	43	5	2.1	5.7	4.7	5.3	0.9	25	P
5	35150	M	63	9.7	7	7	6	2.2	1.1	28	Y
6	23973	F	61	7.2	5.1	11.5	4.4	2.1	1.1	26	Y
7	34278	F	46	5.6	3	5.1	5.7	3.8	1.3	24	Y
8	45703	M	51	12.9	3.9	10.9	3.6	1.1	0.8	29	Y
9	23974	F	60	6.7	6	10.7	4.4	2.1	1.1	26	Y
10	34286	F	45	5.7	3.1	4	5.9	1.8	1.6	24	Y
11	45702	F	54	13	7	7.6	4.9	2.8	0.8	31	Y
12	23975	M	31	4.3	3	12.3	4.1	2.2	0.7	37.2	Y
13	34395	F	60	9.3	5.4	6.8	5.1	2.1	1.1	28	Y
14	43261	F	64	10.2	2	7.9	5.3	3.1	0.9	33	Y
15	23976	F	61	6.9	4.3	12.1	4.4	2	1	29	Y
16	20321	M	50	5.3	4.8	5.9	5.3	1.3	1	19	P

Table 3. The traditional Naïve Bayesian (TNB)

ID	NO_Patient	GENDE R	AGE	Sugar	Urea	HB	Co l	T G	HD L	BM I	Clas s	Prop	c i	Prop_N	Prop_y
73	34498	F	54	14.9	2	13.8	5	12	1.6	26	Y	0.0072	Y	0	0.007207
73	34563	F	57	21.8	4.3	10.2	5.5	1.1	1.4	31	Y	0.0378	Y	0	0.0378
73	34330	M	60	6.7	3.5	7.6	4.7	1.3	0.9	37	Y	0.117	Y	0	0.1171
73	34221	M	26	3.7	4.5	4.9	3.7	1.4	1.1	23	N	0.0508	N	0.50815	3.8E-005
73	34473	M	63	13	6.3	12.4	5	23	0.8	31	Y	0.13035	Y	0	0.13035
73	34409	M	59	10.1	4.2	8.9	5.3	3.1	0.7	27	Y	0.13035	Y	0	0.13035
73	34370	F	56	8.2	1.9	11.7	5.5	5.3	0.9	33	Y	0.0542	Y	0	0.05422
73	34324	F	51	6.6	2.1	5.9	4.1	2.7	1	36	Y	0.0042	Y	0	0.004203
74	34250	M	44	5.1	3.8	5	3.7	1.5	1.1	24	N	0.0608	N	0.060315	3.8E-005
74	34251	F	41	5.1	3.4	4	4.4	1.6	0.8	23	N	0.0211	N	0.021139	1.6E-005
74	34239	F	49	4.9	3.8	4	4.4	0.9	1	23	N	0.0211	N	0.0211	1.6E-005

3.2.2. The MNB algorithm

The use of the traditional naïve Bayesian algorithm was employed as a classifier for diabetes disease. From examining the results of TNB, (22) status errors were found, such as a diabetic patient (Y) classified as a non-diabetic patient (N), shown in Table 4. The overlapping of classifications occurred due to the very nearest probability for two classes. This problem was solved by modifying the naïve Bayesian classifier, the modification was made after the training phase. The modification is summarized in Algorithm 1. Practically, Algorithm 1 is implemented on TrainingDDS with each one of the sets (diabetic, non-diabetic, and predicted-diabetic) separately, and the classification of TestinDDS records has been simultaneously done with the implementation. The Bayes classifier usually works well in MNB, and even when large errors are contained in the probability estimates, excellent classification results can be achieved as shown in Table 5.

Table 4. The TNM module

ID	Sugar	HBAIC	CLASS	PROP	CLASS-TNB	PROP_N	PROP_Y	PROP_P	SUB	THRESHOLD
7	5.6	5.1	Y	0.00029	N	0.00029	1.8e-005	0	0.000272	0.002
10	5.7	4	Y	0.000812	N	0.000812	2e-006	0	0.00081	
31	6.2	4	Y	0.000698	N	0.000698	4.3e-005	0	0.000655	
58	5.7	5	Y	0.00029	N	0.00029	1.8e-005	0	0.000272	
67	6.1	4.9	Y	0.00029	N	0.00029	4.3e-005	0	0.000272	
109	6.1	4.5	Y	0.000698	N	0.000698	4.3e-005	0	0.000655	
168	5.9	4	Y	0.00029	N	0.00029	1.8e-005	0	0.000272	
228	5.6	4.1	Y	0.000698	N	0.000698	4.3e-005	0	0.000655	
351	6.1	4	Y	0.000698	N	0.000698	4.3e-005	0	0.000655	
375	6	4	Y	0.00029	N	0.00029	1.8e-005	0	0.000272	
452	5.8	4	Y	0.000013	N	0.000013	0.000002	0	0.000011	
552	6.1	5.1	Y	0.000338	N	0.000338	3e-006	0	0.000335	
588	5.7	4	Y	0.000129	N	0.000129	8e-006	0	0.000121	
612	6	5	Y	0.00029	N	0.00029	1.8e-005	0	0.000272	
674	5.6	5.2	Y	0.001951	N	0.001951	6e-006	0	0.001945	
694	5.9	4.5	Y	0.000698	N	0.000698	4.3e-005	0	0.000655	
714	5.4	5	Y	0.000698	N	0.000698	4.3e-005	0	0.000655	
717	6	4	Y	0.00029	N	0.00029	1.8e-005	0	0.000272	
725	6	5.5	Y	0.000338	N	0.000338	3e-006	0	0.000335	
750	6.1	4	Y	0.00029	N	0.00029	1.8e-005	0	0.000272	
753	5.7	5.2	Y	0.001951	N	0.001951	6e-006	0	0.001945	
777	6	4	Y	0.000698	N	0.000698	4.3e-005	0	0.000655	

Table 5. The modified Naïve Bayesian

I D	NO_Pati ent	GE N	AG E	Sug ar	URE A	Cr	HB	CH OL	T G	HD L	LD L	VL D	B MI	CLA SS	PROP	C I	PROP _Y	PROP _P
1	34325	M	58	6.7	20.8	80	9.1	6.6	2.9	1.1	4.3	1.3	33	Y	0.0076	Y	0.007	0
2	44335	M	60	10.8	2.1	56	7.6	3.3	1.7	0.9	1.7	0.8	36.6	Y	0.0684	Y	0.068	0
3	23972	F	56	7	4	45	9.2	4.1	0.6	1.3	1.4	0.9	30	Y	0.0487	Y	0.048	0
4	34301	F	43	5	2.1	55	5.7	4.7	5.3	0.9	1.7	2.4	25	P	0.0011	P	0.000	0.001
5	35150	M	63	9.7	7	84	7	6	2.2	1.1	4	1	28	Y	0.1303	Y	0.130	0
6	23973	F	61	7.2	5.1	72	11.5	4.4	2.1	1.1	2.5	0.9	26	Y	0.0698	Y	0.069	0
7	34278	F	46	5.6	3	59	5.1	5.7	3.8	1.3	2.8	1.7	24	Y	0.0002	Y	1.8E-005	0
8	45703	M	51	12.9	3.9	53	10.9	3.6	1.1	0.8	2.3	1	29	Y	0.1171	Y	0.117	0
9	23974	F	60	6.7	6	72	10.7	4.4	2.1	1.1	2.5	0.9	26	Y	0.0698	Y	0.069	0
10	43296	F	45	5.7	3.1	54	4	5.9	1.8	1.6	3.5	0.8	24	Y	0.0008	Y	2e-006	0
11	45702	F	54	13	7	72	7.6	4.9	2.8	0.8	3	1.2	31	Y	0.0698	Y	0.069	0
12	23975	M	31	4.3	3	60	13.3	4.1	2.2	0.7	2.4	15.4	37.2	Y	0.0176	Y	0.017	0
13	34395	F	60	9.3	5.4	47	6.8	5.1	2.1	1.1	3	1	28	Y	0.0542	Y	0.054	0
14	43261	F	64	10.2	2	35	7.9	5.3	3.1	0.9	3.1	1.4	33	Y	0.0542	Y	0.054	0
15	23975	F	61	6.9	4.3	56	12.1	4.4	2.2	1	2.5	0.9	29	Y	0.0698	Y	0.069	0

Algorithm 1. Customized modify Naive Bayesian

Input: TrainingDDS training dataset, S set of eight features, see table (2)

Output: TrainingDDS dataset that has been classified with NB classifier

Steps:

```

For each class Ci in TrainingDDS, compute it as a probability
  For each record R in TrainingDDS, perform steps 3,4,8
    Initialize MaxValue to a small value
    For each class Ci in TrainingDDS, perform steps 5-7
      Compute conditional probability CPi of R at Ci, only for
        values of features of R exists in S, using Eq. (5)
      Compute posteriori probability PPi of R using Eq. (2)
      If PPi greater than MaxValue Then
        MaxValue = PPi and class_label = Ci
        Assign class_label to the class of R
        Apply the modification process
          Detect status error in training (error classifications)
          Sub with absolute the error class from true class.
          Choose the maximum value of subtraction as threshold.
          In testing, check the subtraction value with the threshold
            If it is larger than the threshold,
              then it means true classification
            Else, consider the sugar test attribute
              as a critical factor for classifications.
          End of step.
    End
  End
End

```

End

Note: In steps 1 and 4, i equals: 1 to number of classes in TrainingDDS

4. IMPLEMENTATION AND EXPERIMENTAL RESULTS

This section explains the results according to the standards of evaluation of classifications. In each of these experiments, two classification models were constructed. Subsequently, the constructed models were applied to the same TestingDDS, with the aim of validating the accuracy of the constructed models on the same testing dataset. The classification results are either:

- True positive (TP) answers, which means that positive cases as classified correctly (one of DDS).
- True negative (TN), meaning that negative cases (normal) are correctly classified.
- False positive (FP) answers, meaning that negative cases are wrongly classified as positive (misclassified records as one of DDS).
- False negative (FN) answers, implying the classification of positive cases as negative cases (misclassified records as normal).
- Unknown1 (Predicted- Diabetic true positives Pr-p).
- Unknown2 (Predicted- Diabetic true negatives Pr-n).
- The DR is the ratio between the number of correctly classified records as TP and the total number of intrusion records presented in TestingDDS dataset. It has been computed using;

$$DR = \frac{TP}{TP+FN+Unknown2} \tag{6}$$

- False alarm rate (FAR) is the ratio between numbers of "normal" records classified as an intrusion (FP) and the total number of "normal" records presented in TestingDDS dataset. It has been computed using;

$$FAR = \frac{FP}{TN+FP+Unknown1} \tag{7}$$

- The classification accuracy measures the proportion of correctly classified cases;

$$Acc = \frac{Tp+TN+TPredict}{Tp+Fp+Tn+Fn+unknown} \tag{8}$$

where, Tpredict is (pr-p) and unknown is (Pr-p+Pr-n)

It is noteworthy to mention the following, see Figure 2: Without missing.

- TNB classifier achieved the lowest TP, lowest TN, higher FP, higher FN, lowest unknow1, higher unknown2, and lowest accuracy.
- For MNB classifier, both demonstrated higher TP, higher TN, lowest FP, lowest FN, higher unknown1, lowest unknown2, and higher accuracy.

- The accuracy of the training TNB model has been found to be approximately (63%), while MNB achieved an accuracy of (100%) as shown in Figure 2. It can be deduced that the rate of accuracy of the system DDS for the model MNB is the best average results.

With missing:

- TNB classifier achieved a higher TP, lower TN, higher FP, higher FN, lowest unknow1, higher unknow2 lowest accuracy.
- MNB classifier achieved a higher TP, higher TN, higher FP, higher FN, lowest unknow1, higher unknow2, and higher accuracy.
- The accuracy of the training TNB model has been found to be approximately (62%), while MNB is (71.5%). It can be deduced that the rate of accuracy of the system DDS for the model MNB is the best average results.

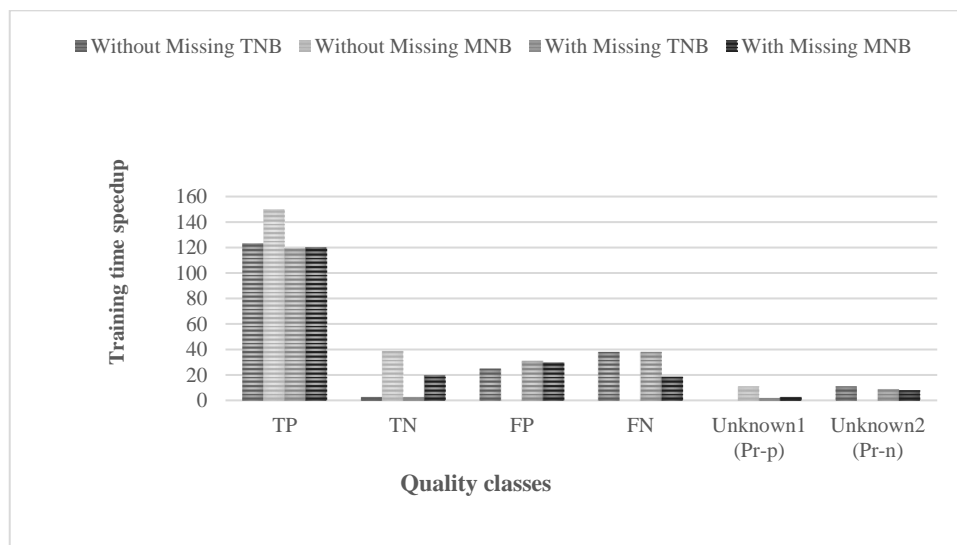


Figure 2. The training dataset results

Table 6 presents the results obtained from the test of the two classifiers with and without missing values; the results are presented as follows, see Figure 3.

Table 6. DR and FAR of two algorithms (TNB and MNB) classifiers testing

Value	Feature	TNB Classifiers		MNB Classifiers	
	Selection Measure	DR	FAR	DR	FAR
Without Missing	Eight	0.715	0.892	1	0
With Missing	Eight	0.716	0.861	0.816	0.566

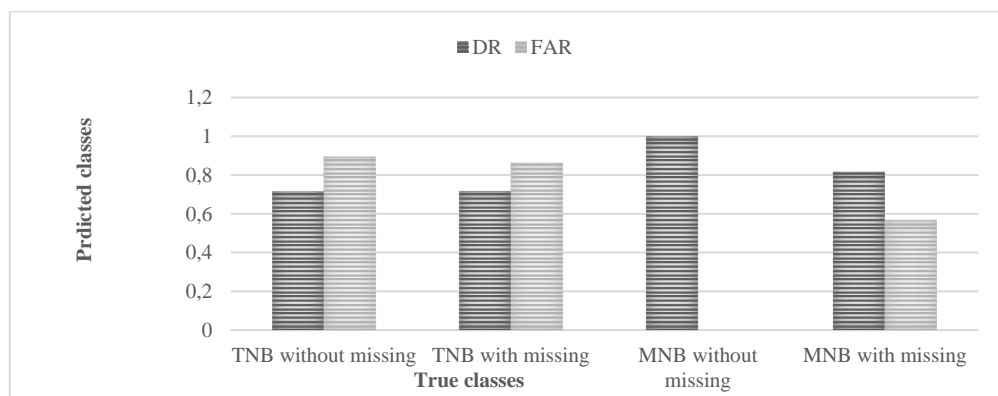


Figure 3. Accuracy of System with missing and without missing

The performance of the classifiers is estimated statistically using the time taken for training. The training time is the time required to build the classifier. The DDS will present TNB as the faster training model, MNB is the second best in terms of speed. The time recorded for TNB was (00:37:99) ms, while that of MNB was (1:16:60) ms. The measure is (Hours: Minutes: Seconds: Milliseconds).

5. CONCLUSION

Diabetes is a good candidate for research since it has a lot of data and a lot of problems. There is a need for a better and more precise method of diagnosing people with that illness. The proposal system proposes a simple modification on the traditional naïve Bayesian algorithm to overcome the overlap that happens in it. MNB gives substantial high and optimal accuracy. It can be observed that a comparison has been done among the algorithms to determine which has the best performance. Despite the fact that the time required by the Tradition Naïve Bayes to build the model is less than that required by others, it was observed that in terms of other parameters, the performance of Modified Naïve Bayes is better. The time needed for building (training time) is estimated to 37:99 ms in the case of TNB while 1:16:60 ms for MNB. When has tested the system in Al- Kindi Hospital and identified that the percentage of error is (4%), and confirmed that the system is working correctly. When taken a large number of patients as input to the system will produce excellent results. Reduced cost of patient management: with such systems, there could be a decrease in the cost required for patient management by sidelining unnecessary investigations and patients follow up. Our future work will be devoted to Applying another classification algorithm, such as support vector machine (SVM) to accomplish a comparison study among the classification algorithms extended the algorithm to deal with more than one. DDS can be uploaded on a free web page to become an online assist readily available to any person who wants to test his/her status. It is a good suggestion to insert automatically the record of the patient who is being chosen to diagnose his/her status, and in this case, a large database can be obtained. Accordingly, the administer can re-train DDS to increase its accuracy and to avoid the care of overlapping.




REFERENCES

- [1] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI Magazine*, vol. 17, no. 3, pp. 37–53, 1996.
- [2] K. N. Kumari and M. G. Subbalakshmi, "A shortest path identification for feature generation and extraction from medicine," *International Journal of Computers Electrical and Advanced Communications Engineering*, vol. 1, no. 3, pp. 11–15, 2012.
- [3] S. Mitra and T. Acharya T, "Data Mining: multimedia, soft computing, and bioinformatics-Google Books," 2005. [Online]. Available: <https://books.google.com.gh/books?hl=en&lr=&id=VPeOaKNfDlGc&oi=fnd&pg=PR7&dq=N> (accessed Sep. 07, 2022).
- [4] R. S. Mans, W. M. P. van der Aalst, and R. J. B. Vanwersch, *Process Mining in Healthcare*. Cham: Springer International Publishing, 2015.
- [5] M. Atzmueller and W. Duijvestijn, "Artificial intelligence," in *30th Benelux Conference, BNAIC 2018*, 2018.
- [6] G. Rong, A. Mendez, E. Bou Assi, B. Zhao, and M. Sawan, "Artificial intelligence in healthcare: review and prediction case studies," *Engineering*, vol. 6, no. 3, pp. 291–301, Mar. 2020, doi: 10.1016/j.eng.2019.08.015.
- [7] M. Matheny, S. T. Israni, and M. Ahmed, "Artificial intelligence in health Care," *National Academy of Medicine*, pp. 1–269, 2018.
- [8] A. Shaheen, "Intelligent decision support system in Diabetic eHealth care from the perspective of elders Asma Shaheen," *Diabetes*, 2009.
- [9] Y. A. Christobel and P. Sivaprakasam, "A new classwise k Nearest Neighbor (CKNN) method for the classification of diabetes dataset," *International Journal of Engineering and Advanced Technology*, vol. 2, no. 3, pp. 396–400, 2013.
- [10] H. Kaur and S. K. Wasan, "Empirical study on applications of data mining techniques in healthcare," *Journal of Computer Science*, vol. 2, no. 2, pp. 194–200, 2006, doi: 10.3844/jcssp.2006.194.200.
- [11] A. Ambica, S. Gandi, and A. Kothalanka, "An efficient expert system for diabetes by Naïve Bayesian classifier," *International Journal of Engineering Trends and Technology (IJETT)*, vol. 4, no. 10, pp. 4634–4639, 2013.
- [12] F. Jiang *et al.*, "Artificial intelligence in healthcare: Past, present and future," *Stroke and Vascular Neurology*, vol. 2, no. 4, pp. 230–243, 2017, doi: 10.1136/svn-2017-000101.
- [13] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nature Medicine*, vol. 25, no. 1, pp. 44–56, 2019, doi: 10.1038/s41591-018-0300-7.
- [14] Y. K. Chan, Y. F. Chen, T. Pham, W. Chang, and M. Y. Hsieh, "Artificial intelligence in medical applications," *Journal of Healthcare Engineering*, vol. 2018, pp. 1–2, Jul. 2018, doi: 10.1155/2018/4827875.
- [15] A. McCallum, "Graphical models, lecture2: bayesian network representation," Retrieved, Lecture, 2019.
- [16] E. Gómez-González *et al.*, "Artificial intelligence in medicine and healthcare: a review and classification of current and near-future applications and their ethical and social impact," *arXiv preprint arXiv:2001.09778*, 2020, [Online]. Available: <http://arxiv.org/abs/2001.09778>.
- [17] G. Sannino, N. Bouguila, G. De Pietro, and A. Celesti, "Artificial intelligence for mobile health data analysis and processing," *Mobile Information Systems*, vol. 2019, Hindawi, 2019, doi: 10.1155/2019/2673463.
- [18] I. R. Ali, A. S. Ahmed, H. K. Tayyeh, H. Kolivand, and M. H. Alkawaz, "Virtual Human for assisted healthcare: application and technology," in *Encyclopedia of Computer Graphics and Games*, Cham: Springer International Publishing, 2019, pp. 1–8.
- [19] T. Lysaght, H. Y. Lim, V. Xafis, and K. Y. Ngiam, "AI-assisted decision-making in healthcare," *Asian Bioethics Review*, vol. 11, no. 3, pp. 299–314, 2019.
- [20] J. C. Bjerring and J. Busch, "Artificial intelligence and patient-centered decision-making," *Philosophy & Technology*, vol. 34, no. 2, pp. 349–371, Jun. 2021, doi: 10.1007/s13347-019-00391-6.




- [21] P. Shah *et al.*, “Artificial intelligence and machine learning in clinical development: a translational perspective,” *npj Digital Medicine*, vol. 2, no. 1, 2019, doi: 10.1038/s41746-019-0148-3.
- [22] J. Awwalu, N. A. Umar, M. S. Ibrahim, and O. F. Nonyelum, “A multinomial Naïve Bayes decision support system for COVID-19 detection,” *FUDMA JOURNAL OF SCIENCES*, vol. 4, no. 2, pp. 704–711, Oct. 2020, doi: 10.33003/fjs-2020-0402-331.
- [23] A. S. Agnal and E. Saraswathi, “Analyzing diabetic data using naive-bayes classifier,” *European Journal of Molecular & Clinical Medicine*, vol. 7, no. 4, pp. 2687–2698, 2020.
- [24] V. Jackins, S. Vimal, M. Kaliappan, and M. Y. Lee, “AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes,” *Journal of Supercomputing*, vol. 77, no. 5, pp. 5198–5219, 2021, doi: 10.1007/s11227-020-03481-x.
- [25] K. Shah, R. Punjabi, P. Shah, and M. Rao, “Real time diabetes prediction using Naïve Bayes classifier on big data of healthcare,” *International Research Journal of Engineering and Technology (IRJET)*, vol. 07, no. 05, pp. 102–107, 2020.

BIOGRAPHIES OF AUTHORS






Jwan Kanaan Alwan    has M.Sc. of information system 2014 from Osmania University in India and currently she is working as lecturer in University of Information Technology and Communications/Baghdad/Iraq. She is interested in pattern classification, medical computing, medical information systems, support vector machines, learning (artificial intelligence). She can be contacted at email: jwanism@uoitc.edu.iq.



Dhulfiqar Saad Jaafar    Date and place of birth: 1987, Baghdad, Iraq. Education: B.Sc. degree in software engineering from College of Al-Rafidain University, 2011 and M.Sc. degree in electric and computer engineering from Altinbas University, 2017. Place of work: Ministry of Education, Baghdad, Iraq. Research interests: machine learning, classification, pattern recognition, linear regression, data mining, deep learning, neural networks, support vector machine. Publications: more than 2 papers and conference proceedings. He can be contacted at email: ghaffoori15@itu.edu.tr, thothosasa@gmail.com.



Itimad Raheem Ali    received the B.Sc. degree in computer science from Al-Nahrain University, Baghdad, Iraq, in 2000, and the M.Sc. degree in computer science also from Al-Nahrain University, Baghdad, Iraq in 2006. She received a Ph.D. degree in computer science from Universiti Teknologi Malaysia, Johor Bahru, Johor. She is currently a lecturer in the Department of Management Information Systems, College of Business Informatics, University of Information Technology and Communications, Baghdad, Iraq. She is research interests include Computer graphics, Facial Animation, Facial Skin color, Emotional Expression of Virtual Image processing and Data Hiding. She has attended international conferences and has published researches in international journals like Elsevier and SPRINGER. She is the editor and chair in many conferences holds in the world from 2016 to yet. He can be contacted at email: itimadra@uoitc.edu.iq.