ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

An Intelligent and Precise Method Used for Detecting Gestational Diabetes in the Early Stages

Safa Abdul Wahab Hameed, Alaa Badeea Ali

Department of Computer Science, Faculty of Engineering and Science, Bayan University, Erbil, Iraq

ABSTRACT

This paper suggests a Naive Bayes classifier technique for identifying and categorizing gestational diabetes mellitus (GDM), GDM is a kind of diabetes mellitus that affects a small proportion of pregnant women but recovers to normal once the baby is born. The Pima Indians Diabetes Dataset was chosen for a comprehensive analysis of this critical and pervasive health disease because it contains 768 patient characteristics acquired from a machine learning source at the University of California, Irvine. The goal of the study is to apply smart technology to categorize diseases with high accuracy and precision, practically free of conceivable and potential faults, to provide satisfying findings. The approach is based on eight major characteristics that are present in the operations that are required to establish a precise and reliable categorization system. This approach involves training and testing on real data, as well as for deciding whether or not to construct a categorization model. The work was compared to earlier work and had a 96% accuracy rating.

Index Terms: Classifier, Feature Selection, Gestational Diabetes, Machine Learning, Naïve Bayes

1. INTRODUCTION

In today's world, diabetes is one of the most frequent diseases [1], whereas diabetes is a non-communicable disease that has a significant impact on people's health today [2]. It is a chronic condition or collection of metabolic diseases in which a person's blood glucose levels remain elevated for an extended period due to insufficient insulin synthesis or improper insulin response by the body's cells [3]. Diabetes mellitus (DM) is a condition that affects more than 60% of the population and has a high mortality rate [4], this disease has increased at an exponential rate in recent years, according to a statistical study published on the World Health Organization's website.

Access this article online

DOI:10.21928/uhdjst.v6n1y2022.pp34-42 E-ISSN: 2521-4217 P-ISSN: 2521-4209

The number of diabetic patients worldwide has increased significantly, from 108 million in 1980 to 422 million in 2014 [5]. A type of diabetes is gestational diabetes. Gestational diabetes is a kind of diabetes that develop during pregnancy [6]. Changes in dietary habits, increased spending power, and climate change, among other factors, are all contributing to an increase in the number of women with gestational diabetes aggravated by pregnancy [7]. Gestational DM (GDM) is a type of glucose intolerance that develops during pregnancy and can cause difficulties for both the mother and the fetus [8]. Women are also more likely to have diabetes-related comorbidities such as renal disease, depression, and poor vision [9]. It can be detected early, which reduces the patient's health risk [10]. However, if the illness is not treated promptly, it can have serious consequences for the kidneys, brain system, retina of the eyes, and heart problems [11], to diagnose diabetes, medical experts require a technique of prediction [12].

Using the PIMA dataset and a variety of machine learning (ML) algorithms, it is feasible to give an advanced technique

Corresponding author's e-mail: Safa Abdul Wahab Hameed, Alaa Badeea Ali, Department of Computer Science, Faculty of Engineering and Science, Bayan University, Erbil, Iraq. E-mail: safa.hamid@bnu.edu.iq; alaa.baban@bnu.edu.iq

Received: 07-12-2021 Accepted: 18-02-2022 Published: 20-03-2022

for diabetes prediction. Because deep learning algorithms may be used in a variety of ways in this industry, Models based on Artificial Neural Networks (ANN) and the Quasi-Newton technique, for example [13]. This area lends itself well to ML methods. Models are trained using ML techniques. There are three types of ML algorithms: Supervised learning (in which datasets are labeled and Regression and Classification techniques are used), unsupervised learning (in which datasets are not labeled and techniques such as dimensionality reduction and clustering are used), and reinforcement learning (in which the model learns from its every action). ML is a rapidly developing new technology with several applications [14]. ML techniques have advanced at a breakneck pace and are now widely used in a variety of medical applications [15], one of the most commonly explored challenges by DM and ML researchers is classification [16]. One of the most crucial parts of supervised learning is classification. Picking the proper classification model is a trade-off between performance, the execution time of models, and scalability. Parameter adjustment should also be considered in order to improve model performance. ML training data are an important input to an algorithm that comprehends and memorizes information from such data to predict the future. Understanding the significance of the training set in ML can assist you in obtaining the appropriate quality and amount of training data for your model training. Once you understand why it's essential and how it influences model prediction, you'll be able to select the best method based on the availability and compatibility of your training data set [17]. It is vital to develop predictive algorithms that are both accurate and simple to use when evaluating large amounts of data and converting it into useful information [18] ML methods are commonly utilized for detection and classification [19]. This diagnosis allows for proper treatment to begin as soon as feasible, avoiding deaths [20], [21]. People will be able to seek treatment for this condition if it can be detected and predicted at an early stage [22]. This type of disease is the gestational diabetes is the focus of this research, in this research an effective method was used, which is a Naïve Bayes classifier, it used to detect and identify gestational diabetes and give high performance and accurate results.

2. LITERATURE REVIEW

Similar works on diabetes analysis, prediction, and diagnosis are reviewed in this section. It uses a variety of classification and ML algorithms to handle diabetes management prediction problems.

In Pradhana *et al.* [5], the suggested methodology analyzed publically accessible data collected from diabetic patients to identify the causes of diabetes, the most afflicted age groups, job styles, and eating patterns. ANN are used in the model to detect diabetes and determine its kind. The authors utilized the "Pima Indian Diabetes" dataset, which has the maximum accuracy of 85.09**%**. 768 patients' medical histories are included in the dataset. Where as in the Filho *et al.* [7], the suggested method improved the accuracy of the classification techniques by focusing on identifying the features that fail in early diagnosis of Diabetes Miletus utilizing Predictive analysis using Support vector machine (SVM) and Naïve Base (NB) algorithms. The accuracy of the improved SVM is 77**%**, while the accuracy of the NB is 82.30**%**. In Prasanth *et al.* [2] explained, the captured data were fed into supervised ML techniques. The Pima Indians Diabetes Dataset was utilized in this study, and a model was created using SVM, CatBoost, and Relative frequency to predict DM, with an accuracy of 86.15**%**. And in Rawat and Suryakant [18], a comparison was done between suggested approaches and previously published studies. In this work, five ML algorithms, AdaBoost, LogicBoost, RobustBoost, Naive Bayes, and Bagging, were proposed for the analysis and prediction of DM patients. The suggested methodologies were tested on a data set of Pima Indians with diabetes, and the proposed algorithm, Bagging, was applied on the same database with an accuracy of 81.77**%**.

The two algorithms Naive Bayes and SVM used in Gupta *et al.* [1] as classification models, and feature selection to improve the model's accuracy. The accuracy, precision, and recall values were used to evaluate the results. The model's improved performance was calculated using the k-fold cross-validation technique. In Rajivkannan and Aparna [22], the aim of this study was to build an objective method to evaluate DM risk from past GDM data recorded 15 years ago and find a shortlist of the most informative indicators. The research steps involve pre-processing data to evaluate missing values MVs, finding the most informative attributes, and testing standard classification algorithms to combine into the most effective voting meta-algorithm. Meta-algorithm-based classification of limited anamnestic GDM related data for DM prediction is proving. Relative frequency of occurrence (RFO) analysis of attributes combined with voting meta-algorithm helped find the optimal amount of attributes giving the best possible classification result. The algorithm applied to two-class data set with 12 selected attributes produced an accuracy of 75.85. In [17], the major goal of this research is to look at different forms of machine-learning classification algorithms and compared them. In this study, use machine-learning classification algorithms to detect the start of diabetes in diabetic patients. The top performing algorithm, Logistic Regression, has an accuracy of 80%. In Moon [21], the research provided a biological ML

method that is both efficient and effective which is applied on Pima Indian diabetic database (PIDD). The proposed ensemble of SVM and back-propagation neural network (BP-NN) tested on diabetes diagnosis; one of the most frequently investigated topics in bioinformatics. The findings reveal an accuracy of 88.04% on this problem. In Sanakal and Jayakumari [16], the application of Fuzzy C-means clustering (FCM), FCM, and SVM on a collection of medical data linked to diabetes diagnostic difficulties was the subject of this work. The medical data comprises nine input variables relating to diabetes clinical diagnosis and one output attribute that indicates whether or not the patient has been diagnosed with diabetes. FCM had the best outcome, with an accuracy of 94.3%. In Lavanya and Rani [23], this work provided a quick overview of the old and new data mining approaches utilized in diabetes. This study used Fcm and Svm procedures and got an accuracy of 94.3%. In Sarwar *et al.* [4], the results showed that the ensemble approach had a 98.60% accuracy, which combines the predictive performance of numerous AI-based algorithms and is superior to all other individual competitors. ANN, Naive Bayes, SVM, and K-Nearest Neighbor are the techniques that were more precise than the others (K-NN). About 400 persons were included in the database, which came from all across the world. And in Resti *et al.* [9] a model validation built which based on 5-fold cross-validation, which divided the data into training and test data. The Gaussian Nave Bayes was the best strategy for predicting diabetes diagnosis, according to the model validation results. The contribution of this research was that the Multinomial Naïve Bayes method's performance metrics all exceed 93%. With the same explanatory factors, these findings are useful in predicting diabetes status. In this paper, the proposed method is implemented on Pima Indians Diabetes Dataset. Here in this paper, the work was presented using a Naive Base algorithm. The work was carried out in stages that relied on training and testing on real data based on certain characteristics as explained later, and a classification result was obtained with high efficiency and accuracy.

3. METHOD

This work is done to diagnose gestational diabetes by classification using a Naive Base algorithm. This work includes several stages, including training and testing on real data, to adopt and use the system. Following the stages to implement the method:

3.1. System Diabetic Detection

The Diabetes Classification System includes two phases; the first is the training stage, which has particular functions such as reading the diabetes dataset, feature selection, discretization, and the classifier model used to create the decision rules. The second stage is the testing stage, which includes the following particular functions: read data set, discretization, decision rules, and output [1], [24], as illustrated in Fig. 1.

3.2. Data Set

The data set has been taken from 768 women, (500 negative cases and 268 positive cases) from 21 years old and above, and eight recorded features as follows:

- Number of previous pregnancies
- Plasma glucose concentration at 2-h in an oral glucose tolerance test
- Diastolic blood pressure (mm Hg)
- Triceps skinfold thickness (mm)
- 2-H serum insulin (mu U/ml),
- Body mass index (weight in kg/[height in m]^2)
- Diabetes pedigree function
- Age (years)

There are a variety of causes for incomplete data, including patient death, device problems, and respondents' reluctance to answer particular questions.

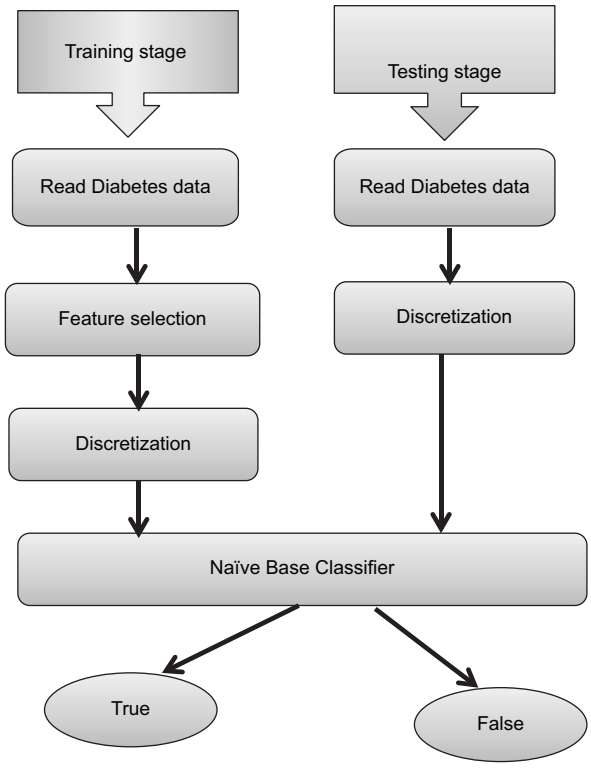

Fig. 1. The structure of the training and testing stages of the diabetes classification system.

Every patient in the database is a Pima Indian lady of at least 21 years of age who lives in or around Arizona. There are eight properties in each dataset sample (attributes), as shown in Table 1.

A sample of the data used is shown in Table 2: The row header in the table related to the column header (features) in Table 1.

There are 768 samples in all, divided into two groups. The following is the distribution of classes:
1. Normal group: (500) samples
2. Abnormal group: (268) samples.

Fold creation: The complete labeled dataset is separated into mutually exclusive folds to perform the cross validation procedure. The 768 cases in this study were chosen from the PIDD database, which means:
1. The training-set contains 499 occurrences, 325 of which are normal and 174 of which are aberrant (i.e., 65%)
2. There are 269 examples used in the testing, 175 of which are normal and 94 of which are abnormal (i.e., 35%) as shown in the Table 3.

3.3. Feature Selection Stage

The PIDD data set has eight attributes in each sample. According to the entropy of the property with class, Let S (training data) be a data-set of C outputs. For the classification issue with C classes, let P (I) indicate the proportion of S that belongs to Class I, where I is distinct from one to C [1], [23], [24].

p is the simple diversity index (I) (1)

The information–theoretical approach of assessing the quality of the split is entropy. It calculates the quantity of information in a given attribute.

$$\text{Entropy}(S) = \sum_{I=1} (-P(I)\,P(I)) \quad (2)$$

The information gain of the example set S on the attribute A is defined as Gain(S,A).

$$\text{Gain}(S,A) = \text{Entropy}(S) - \left(\left(\frac{|SV|}{|S|}\right) \times \text{Entropy}(SV)\right) \quad (3)$$

Where SV = subset of S in which feature A has value V, |SV| = amount of data in SV, and |S| = amount of elements in S, and is over every value V of every conceivable value of the attribute A, a sub-set property is chosen from eight options. It takes a long time to choose more than five properties. Selecting <5 attributes, on the other hand, make the Diabetes Classification System less time-consuming but less accurate. According to the greater entropy, the ideal five attributes (1, 3, 4, 5, and 7) for describing diabetes have been determined and will be used as an input to the classification step.

TABLE 1: Number of Feature of PIDD

Feature	Description	Range
1	Number of times pregnant	1-4
2	Plasma glucose concentration a 2 h in an oral glucose tolerance test.	I. 120-140
3	Diastolic blood pressure (mm Hg).	80-90
4	Triceps skin fold thickness (mm).	12 mm (male)-23(female)
5	2-h serum insulin (IU/ml).	16-166 mlu/L
6	Body mass index (weight in kg/(height in m)^2)	18-24.9 kg/m2
7	Diabetes pedigree function Feature 8: 2-h serum insulin (IU/ml).	1-3
8	Age (years).	21 and above

TABLE 2: A sample of PIMA Indian diabetes dataset

1	2	3	4	5	6	7	8	Class
1	116	78	29	180	36.1	0.496	25	Normal
2	130	96	0	0	22.6	0.268	21	Normal
0	129	110	46	130	67.1	0.319	26	Abnormal
0	135	68	42	250	42.3	0.365	24	Abnormal
4	112	78	40	0	39.4	0.236	38	Normal

TABLE 3: Number of cases tested and trained

Training	No. Of cases	Normal	Abnormal
	499	325	174
Testing	269	175	94

3.4. Discretization

Before the categorization procedure in the Diabetes Classification System, a crucial phase must be completed. It is necessary to transform the numerical values of eight attributes to category values. This is accomplished by splitting the range of values for eight characteristics into k equal-sized bins, or equal width intervals, where k is a user-selected quantity based on the length of data. Where it can be done in a mechanism of the Equal Width Interval Discretization (EWID) algorithm, by the method used in this algorithm,

```
Algorithm (1): Equal Width Interval Discretization (EWID)

Input: Eight attributes have numerical values.

Output: Eight attributes have categorical values.

Begin

minimum1 = Op1(0) : maximum1 = Op 1(0) : minimum 2 = Op 2(0) : maximum 2 = Op 2(0) : minimum 3 = Op 3(0) : maximum 3 = Op 3(0)

minimum4 = Op 4(0) : max4 = Op 4(0) : minimum 5 = Op 5(0) : maximum 5 = Op 5(0) : minimum 6 = Op 6(0) : maximum 6 = Op 6 (0)

minimum 7 = Op 7(0) : maximum 7 = Op 7(0) : minimum 8 = Op 8(0) : maximum 8 = Op 8(0)


For x = 1 To Len of data set {

If Op1(x) > maximum1 Then maximum1 = Op1(x) : If Op1(x) < minimum 1 Then minimum 1 = Op1(x)

If Op2(x) > maximum2 Then maximum2 = Op2(x) : If Op2(x) < minimum 2 Then minimum 2 = Op2(x)

If Op3(x) > maximum3 Then maximum3 = Op3(x) : If Op3(x) < minimum 3 Then minimum 3 = Op 3(x)

If Op4(x) > maximum4 Then maximum4 = Op4(x) : If Op4(x) < minimum 4 Then minimum 4 = Op4(x)

If Op5(x) > maximum5 Then maximum5 = Op5(x) : If Op5(x) < minimum 5 Then minimum 5 = Op5(x)

If Op6(x) > maximum6 Then maximum6 = Op6(x) : If Op6(x) < minimum 6 Then minimum 6 = Op6(x)

If Op7(x) > maximum7 Then maximum7 = Op7(x) : If Op7(x) < minimum7 Then minimum 7 = Op7(x)

If Op8(x) > maximum8 Then maximum8 = Op8(x) : If Op8(x) < minimum8 Then minimum8 = Op8(x) }

K=3 // possibilities number

G1 = Round(((maximum1 - minimum1) / k), 3) : G2 = Round(((maximum2 - minimum2)/ k), 3)

G3 = Round(((maximum3 - minimum3) / k), 3) : G4 = Round(((maximum4 - minimum4)/ k), 3)


G5 = Round(((maximum5 - minimum5) / k), 3) : G6 = Round(((maximum6 - minimum6)/ k), 3)

G7 = Round(((maximum7 - minimum7) / k), 3) : G8 = Round(((maximum8 - minimum8)/ k), 3)

// Identify k ranges for each of the eight attributes.


LowOp1 = (minimum1 + G1) : medOp1 = (minimum 1 + (2 * G1)) : highOp1 = (minimum1 + (3 * G1))

LowOp2 = (minimum2 + G2) : medOp2 = (minimum2 + (2 * G2)) : highOp2 = (minimum2 + (3 * G2))

LowOp3 = (minimum3 + G3) : medOp3 = (minimum3 + (2 * G3)) : highOp3 = (minimum3 + (3 * G3))

LowOp4 = (minimum4 + G4) : medOp4 = (minimum4 + (2 * G4)) : highOp4 = (minimum4 + (3 * G4))

LowOp5 = (minimum5 + G5) : medOp5 = (minimum5 + (2 * G5)) : highOp5 = (minimum5 + (3 * G5))

LowOp6 = (minimum6 + G6) : medOp6 = (minimum6 + (2 * G6)) : highOp6 = (minimum6 + (3 * G6))

LowOp7 = (minimum7 + G7) : medOp7 = (minimum7 + (2 * G7)) : highOp7 = (minimum7 + (3 * G7))

LowOp8 = (minimum8 + G8) : medOp8 = (minimum8 + (2 * G8)) : highOp8 = (minimum8 + (3 * G8))

End
```

Fig. 2. The structure for the EWID algorithm code.

we have obtained the traits under a specified range. It is an important step before classifying to convert it into categories, to be used for training. [25], [26] [27], which consists of steps as shown in Fig. 2.

3.5. Classifier Model

The most well-known task classification is the constructing classifier model. This structure is used to determine the diabetes class, which can be either normal or abnormal. The Diabetes (Training-Set) database is made up of attribute-value representations for a large number of patients, with five categorical characteristics (1, 3, 4, 5, and 7) and class attributes. Those characteristics are fed into the learning classifier model. The classifier model is used to predict the new case. The decision is used to build the classifier model in this work, which is based on training diabetes [1], [25]. This mechanism is implemented using a Naive base algorithm, where the algorithm is fed by the characteristics from the training set, and this helps in building a classifier model for prediction based on the decision, the steps of this algorithm and the mechanism used in it are explained as shown in Fig. 3.

4. RESULTS

As previously stated, the diabetes categorization system's initial data included Pima Indian diabetes illness measures. Before classification, the retrieved numerical values from attributes must be transformed into categorical values, which will be used to train the classifier using the EWID method.

Table 4 shows the categorical values of the five attributes according to the Diabetes Classification System. Attributes, attribute-value, and range of values are the three fields that

```
Algorithm (2): Naïve Base Algorithm
Output: decision
Begin
For i = 0 To 1 // two group
Q = 0
For j = 0 To LenTr
If Opr(i) = Op(j) Then Q = Q + 1
End for
Opr (i) = Math.Round((Q / CT), 3)
end for
For G = 1 To 8
If G = 1 Then a = b1 : If G = 2 Then a = b2 : If G = 3
Then a = b3
If G = 4 Then a = b4 : If G = 5 Then a = b5 : If G = 6
Then a = b6
If G = 7 Then a = b7 : If G = 8 Then a = b8
For i = 0 To a
For k = 0 To 1
Q = 0 :Y = 0
For j = 0 To LenTr
If G = 1 Then If P1(i) = F1(j) And Opr (k) = Op(j) Then Q=
Q+ 1
If G = 2 Then If P2(i) = F2(j) And Opr(k) = Op(j) Then Q =
Q + 1
If G = 3 Then If P3(i) = F3(j) And Opr(k) = Op(j) Then Q =
Q + 1
If G = 4 Then If P4(i) = F4(j) And Opr(k) = Op(j) Then Q =
Q + 1
If G = 5 Then If P5(i) = F5(j) And Opr(k) = Op(j) Then Q =
Q + 1
If G = 6 Then If P6(i) = F6(j) And Opr(k) = Op(j) Then Q =
Q + 1
If G = 7 Then If P7(i) = F7(j) And Opr(k) = Op(j) Then Q =
Q + 1
If G = 8 Then If P8(i) = F8(j) And Opr(k) = Op(j) Then Q =
Q + 1
If G = 9 Then If P9(i) = F9(j) And Opr(k) = Op(j) Then Q =
Q + 1
If G = 10 Then If P 10(i) = F 10(j) And Opr(k) = fc(j) Then
Q = Q + 1
Next
// likelihood
If G = 1 Then pp1(i, k) = Math.Round((Q / (QP(k) *CT)), 3)
If G = 2 Then pp2(i, k) = Math.Round((Q / (QP(k) *CT)), 3)
If G = 3 Then pp3(i, k) = Math.Round((Q / (QP(k) *CT)), 3)
If G = 4 Then pp4(i, k) = Math.Round((Q / (QP(k) *CT)), 3)
If G = 5 Then pp5(i, k) = Math.Round((Q / (QP(k) *CT)), 3)
If G = 6 Then pp6(i, k) = Math.Round((Q / (QP(k) *CT)), 3)
If G = 7 Then pp7(i, k) = Math.Round((Q / (QP(k) *CT)), 3)
If G = 8 Then pp8(i, k) = Math.Round((Q / (QP(k) *CT)), 3)
End for
End for
End for
End
```

Fig. 3. Naïve Base Algorithm.

make up this table. The five attributes are shown in the first field (How many times pregnant, Diastolic blood pressure, Triceps skin fold thickness, serum Insulin and diabetes pedigree function Feature). The categorical values of five attributes are shown in the second field. The range of values obtained by the EWID method is represented in the third field.

Table 5 shows the categorical values samples that were obtained by changing numerical property values of the diabetes case with the class attribute according to the range of value field in Table 4.

Table 6 shows the results of the top five attributes chosen for use in the Classifier model. The following properties were used to train and evaluate the classifier model: Preg signifies the number of pregnancies a woman has had, Pres the Diastolic blood pressure, Skin the thickness of the Triceps skin folds, Insu the serum insulin, and Pedi the Diabetes pedigree function feature. The entropy for each feature is calculated using the entropy equation (2).

For the diabetes instances indicated in the tables, the following naïve Bayés classifier was trained: The likelihood of the number of times pregnant features for the three ranges of two classes (normal and abnormal) is represented in Table 7.

The normal class is represented by the number 0.322, while the abnormal class of the low range is represented by the value 0.551. The abnormal class of medium-range is represented by the value 0.301. The number 0.241 represents the typical normal class, whereas the value 0.148 represents the aberrant high-range class. The likelihood of each class being equal to one is shown by the end row.

The likelihood of the number of skin characteristics for the three ranges of two classes (normal and abnormal) is represented in Table 8. The number 0.35 represents the normal class, whereas the value 0.431 represents the low-range abnormal class. The values 0.276 and 0.179 represent the normal and abnormal classes in the medium range, respectively, while the values 0.403 and 0.39 represent the normal and abnormal classes in the high range, respectively. The likelihood of each class being equal to one is shown by the end row.

The likelihood of the number of ins features for the three ranges of two classes is represented in Table 9. (Normal, abnormal) The values 0.977 and 0.865 reflect the normal and abnormal classes, respectively, of the normal –range. The normal class is represented by the number 0.023, whereas the abnormal class is represented by the value 0.006.

The likelihood of the number of pedi characteristics for the three ranges of two classes (abnormal and normal) is shown by Table 10, where the value 0.874 represents the normal class and the value 0.935 represents the abnormal class of the low –range. The number 0.126 represents the normal class, whereas the value 0.065 represents the high-range abnormal class. The probability of each class that is equal to one is summed in the last row.

Table 11 provides the confusion matrix of classifier implementation retrieved from the testing stage using NB.

The accuracy and error rate for diagnosed cases are calculated using Table 11. The values used in calculating the accuracy of the NB classifier using the accuracy equation (4),

Where the:

TP: true positive

TN: true negative

TP and TN are added together, then divided by the sum of all with FP (false positive) and FN (false negative), and the error is computed using the equation (5) [28].

$$\begin{aligned} Accuracy &= (TP+TN) \setminus (TP+TN+FP+FN) \\ Accuracy &= (173+87)/(173+4+5+87) \\ &= 0.96 \end{aligned} \quad (4)$$

$$Error = (FP+TN)/(TP+TN+FP+FN) = 0.334 \quad (5)$$

TABLE 4: Categorical features

Attributes	Attribute-value	Range of Values
Number of times pregnant	Low	(0–2)
	Medium	(3–5)
	High	(6–17)
Diastolic blood pressure	Low	(0–80)
	Medium	(80–100)
	High	(100–122)
Triceps Skin fold thickness	Low	(0–20)
	Medium	(20–60)
	High	(60–99)
Serum insulin	Normal	(0–280)
	Abnormal	(280–860)
Diabetes pedigree	Low	(0.084–1.251)
	High	(1.251–2.42)

TABLE 5: Samples of categorical features values

Id	Preg	Press	Skin	Insu	Pedi	Class
1	Low	Low	Low	Normal	Low	Normal
2	Low	Medium	Medium	Normal	High	Normal
3	High	Low	Medium	Normal	Low	Normal
4	High	Medium	Medium	Abnormal	High	Abnormal
5	Low	High	Medium	Normal	High	Abnormal
6	Medium	High	High	Abnormal	High	Abnormal

TABLE 6: Entropy of categorical features values

Entropy	Preg	Plas	Skin	Skin	Ins	Mass	Pedi	Age
	0.71	0.13	0.49	0.88	0.55	0.02	0.64	0.08

TABLE 7: Preg Feature probability

Preg	Normal	Abnormal
Low	0.322	0.551
Medium	0.437	0.391
High	0.241	0.148
SUM	1	1

TABLE 8: Skin feature probability

Preg	Normal	Abnormal
Low	0.35	0.431
Medium	0.276	0.179
High	0.403	0.39
SUM	1	1

TABLE 9: Ins Feature probability

Preg	Normal	Abnormal
Normal	0.977	0.865
Abnormal	0.023	0.006
SUM	1	1

TABLE 10: Pedi feature probability

Preg	Normal	Abnormal
Low	0.874	0.935
High	0.126	0.065
SUM	1	1

TABLE 11: The confusion matrix using Naïve Bayes classifier

Predicate Class	Actual Class	
	Normal	Abnormal
Normal	(TP) 173	(FP) 4
Abnormal	(FN) 5	(TN) 87

These values are compensated according to the work performed and the results obtained.

This proposed work was used to enhance accuracy using the Naive Base classifier method, where the mechanism was implemented with a performance that gives higher accuracy than the accuracy obtained from previous studies. This work has been compared with the previous works as shown in Table 12, as each work used the appropriate mechanism for diagnosing and classifying this disease and obtained an appropriate accuracy rate for the work. After studying and analyzing this problem, a high percentage of accuracy was obtained.

5. CONCLUSION

In this study, the NB classifier was used, we attempted to provide an approach for identifying the classification method for detecting and classifying diabetes at an early stage. There are eight properties in each dataset sample (attributes), divided into two classes normal and abnormal, and used in two stages (training and testing), during the training stage, specific functions were performed, such as reading the diabetes dataset, feature selection, discretization, and the classifier model used to create the decision rules, and during the testing stage, specific functions were performed, such as reading the diabetes dataset, discretization, decision rules, and output. The findings of the experiments were run on the dataset and compared with the previous works, and a system that can

TABLE 12: Comparison of accuracy with previous works

Authors	The year	dataset	The method used	Accuracy
Sneha N. and Gangil T	2019	Pima Indians Diabetes Dataset	1) Support vector machine SVM 2) Naïve Base NB	77% 82.30%
Islam M. A. and Jahan N.	2017	Pima Indians Diabetes Dataset	Logistic Regression algorithm	80%
Rawat V and Suryakant S.	2019	Pima Indians Diabetes Dataset	Bagging method	81.77%
Pradhana N , Rania G, Singh V, Dhaka V. S and Pooniab R. C.	2020	Pima Indian Diabetes dataset	Artificial neural networks	85.09%
Prasanth S, Banujan K and Btgs K	2021	Pima Indians Diabetes Dataset	The adaptation model of SVM, CatBoost, and Random Forest (RF)	86.15%.
Zolfaghar	2012	Pima Indians Diabetes Dataset	the support vector machine (SVM) and back-propagation neural network (BP-NN)	88%
Resti Y., Kresnawati E. S., Dewi N. R., Zayanti D. A. and Eliyati N.	2021	Pima Indians Diabetes Dataset	Naive Bayes, Discriminant Analysis, and Logistic Regression	93%
Sanakal R. and Jayakumari S. T.,	2014	Pima Indians Diabetes Dataset	Fuzzy C Means Clustering	94.3&
Jayanthi N. , Babu .V. B. and Rao S.	2016	Pima Indians Diabetes Dataset	FCM and SVM	94.3%
Hameed S. A. and Baban A. B. (The proposed method in this paper)	2021	Pima Indians Diabetes Dataset	Naive Base classifier	96%

reliably diagnose and categorize gestational diabetes was shown to be 96% accurate.

6. REFERENCES

[1] S. Gupta, H. K. Verma and D. Bhardwaj. "Classification of diabetes using naïve Bayes and support vector machine as a technique". *Operations Management and Systems Engineering*, pp. 365-376, 2020.

[2] S. Prasanth, K. Banujan and K. Btgs. "*Hyper Parameter Tuned Ensemble Approach for Gestational Diabetes Prediction*". International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT), IEEE. pp. 18-23, 2021.

[3] N. Sneha and T. Gangil. "Analysis of diabetes mellitus for early prediction using optimal features selection". *Journal of Big Data*, vol. 6, p. 13, 2019.

[4] A. Sarwar, M. Ali, J. Manhas and V. Sharma. "Diagnosis of diabetes Type-II using hybrid machine learning based ensemble model". *International Journal of Information Technology*, vol. 12, pp. 419-428, 2020.

[5] N. Pradhana, G. Rania, V. Singh, V. S. Dhaka and R. C. Pooniab. "Diabetes prediction using artificial neural network". In: *Deep Learning Techniques for Biomedical and Health Informatics*, ScienceDirect, pp. 327-339, 2020.

[6] M. D. Okpor. "Prognostic diagnosis of gestational diabetes utilizing fuzzy classifier". *International Journal of Computer Science and Network Security*, vol. 15, no. 6, pp. 44-48, 2015.

[7] E. G. Filho, P. R. Pinheiro, M. C. D. Pinheiro, L. C. Nunes and L. B. G. Gom. "Heterogeneous methodology to support the early diagnosis of gestational diabetes. *IEEE Access*, vol. 99, p. 1, 2019.

[8] M. Marozas, S. Sosunkevič, M. Francaitė-Daugėlienė, D. Veličkienė and A. Lukoševičius. "Algorithm for diabetes risk evaluation from past gestational diabetes data". *Technology and Health Care*, vol, 26, no. 4, pp. 637-648, 2018.

[9] Y. Resti, E. S. Kresnawati, N. R. Dewi, D. A. Zayanti and N. Eliyati. Diagnosis of diabetes mellitus in women of reproductive age using the prediction methods of naive bayes, discriminant analysis, and logistic regression. *Science and Technology Indonesia*, vol. 6, no. 2, pp. 96-104, 2021.

[10] M.A. Islam and N. Jahan. Prediction of Onset Diabetes using Machine Learning Techniques. *International Journal of Computer Applications*, vol. 180, no. 5, pp. 7-11, 2017.

[11] R. Saxena, S. K. Sharma and M. Gupta. "Analysis of machine learning algorithms in diabetes mellitus prediction". Journal of Physics: Conference Series, vol. 1921, p. 012073, 2021.

[12] N. Jayanthi, V. B. Babu and S. Rao. "Data mining techniques for CPD of diabetes". *International Journal of Engineering Computational Research and Technology*, 2014.

[13] K. Lakhwani, S. Bhargava, K. K. Hiran, M. M. Bundele and D. Somwanshi. "*Prediction of the Onset of Diabetes Using Artificial Neural Network and Pima Indians Diabetes Dataset*". 5th IEEE International Conference on Recent Advances and Innovations in Engineering, pp. 1-6, 2020.

[14] R. Zolfaghar. "Diagnosis of diabetes in female population of pima indian heritage with ensemble of BP neural network and SVM". *International Journal of Computational Engineering and Management*, vol. 15, no. 4, pp. 115-121, 2012.

[15] A. Kaushik, A. Sehgal, S. Vora, V. Palan and S. Patil. "*Presaging The Signs Of Diabetes Using Machine Learning Algorithms*". 12th International Conference on Computing Communication and Networking Technologies, 2021.

[16] R. Sanakal and S. T. Jayakumari. "Prognosis of diabetes using data mining approach-Fuzzy C Means clustering and support vector machine". *International Journal of Computer Trends and Technology*, vol. 11, no. 2, pp. 94-98, 2014.

[17] H. Naz and S. Ahuja. "Deep learning approach for diabetes prediction using PIMA Indian dataset". *Journal of Diabetes and Metabolic Disorders*, vol. 19, no. 1, pp. 391-403, 2020.

[18] V. Rawat and S. Suryakant. "A classification system for diabetic patients with machine learning techniques". *International Journal of Mathematical, Engineering and Management Sciences*, vol. 4, no. 3, pp. 729-744, 2019.

[19] P. Kaur and R. Kaur. "Comparative analysis of classification techniques for diagnosis of diabetes". In: Jain, L., Virvou, M., Piuri, V. and Balas, V. (eds.), *Advances in Bioinformatics, Multimedia, and Electronics Circuits and Signals Advances in Intelligent Systems and Computing*. Vol. 1064. Springer, Singapore, 2020.

[20] L. Jonk. "*Chronic Disease Prevention a Vital Investment*". World Health Organization, Geneva, Switzerland, 2005.

[21] L. Moon. "*Prevention of Cardiovascular Disease, Diabetes and Chronic Kidney Disease: Targeting Risk Factors*". Vol. 118. AIHW, 2009. Available from: http://www.aihw.gov.au/publications/index.cfm. [Last accessed on 2022 Mar 09].

[22] A. Rajivkannan and K. S. Aparna. "A survey on diabetes prediction using machine learning techniques". *International Journal of Research in Engineering, Science and Management*, vol. 4, no. 11, pp. 51-54, 2021.

[23] D. Lavanya and K. U. Rani. "Performance evaluation of decision tree classifiers on medical datasets". *International Journal of Computer Applications*, vol. 26, no. 4, pp. 1-4, 2011.

[24] R. Raja, I. Mukherjee and B. K. Sarkar. "A machine learning-based prediction model for preterm birth in Rural India". *Journal of Healthcare Engineering*, vol. 2021, p. 6665573, 2021.

[25] A. Saleha and F. Nasari. "Implementation of equal-width interval discretization in naive bayes method for increasing accuracy of students' majors prediction". Lontar Komputer Jurnal Ilmiah Teknologi Informasi. Vol. 9, no. 2, pp. 104-113, 2018.

[26] R. Dash, R. L. Paramguru and R. Dash. "Comparative analysis of supervised and unsupervised discretization techniques". *International Journal of Advances in Science and Technology*, vol. 2, no. 3, pp. 29-37, 2011.

[27] J. Dougherty, R. Kohavi and M. Sahami. "*Supervised and Unsupervised Discretization of Continuous Features*". In: Proceedings of the Twelfth International Conference on International Conference on Machine Learning (ICML'95). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1995, pp. 194-202.

[28] J. Han and M. Kambar. "*Data Mining: Concepts and Techniques*". 2nd ed. Morgan Kaufmann Publisher, Burlington, Massachusetts, 2006.