دوازدهمین کنگره ملی سراسری
فناوریهای نوین درحوزه توسعه پایدار ایران
12th National Congress of
the New Technologies in Sustainable Development of Iran
senaconf.ir

Diabetes Diagnosis from Big Data using Fuzzy-Neural Chaotic Tree

Banafsheh Saleh
Department of Information Technology
Sabzevar Islamic Azad University
Sabzevar, Iran
banafshehsaleh07@gmail.com

Hesam Hasanpour
Department of Computer
Sabzevar Branch, Islamic Azad University
Sabzevar, Iran
hesam_78@yahoo.com

Abstract—**Today, diabetes is considered an important disease in the world. The statistics show that this disease is developing worldwide. For this reason, smart and automated systems are considered as a challenge in the medical world. The methods available in computer science meet this need. To date, various methods have been proposed to diagnose and predict diabetes, but there are some errors that researchers are seeking to resolve. Data mining is used as a technical science in identifying and extracting new knowledge of data. In this research, a new method has been developed for categorizing diabetic data, which consists of three parts: the first part is the preprocessing, in which data normalization operations are performed, then the extraction and selection of attributes. Finally, classification operations is done based on the principles of data mining. The classification results can be used to predict diabetes in different individuals. The use of criteria such as sensitivity, specificity, accuracy is used to evaluate the results. The proposed approach is based on the combination of fuzzy-neural in chaotic mode along with K-means tree. The results indicate that the proposed method is more suitable than the previous methods.**

Keywords- Diabetes Detection, Data Mining, Fuzzy-Neural, Chaos Theory, K-means Tree

1. Introduction

Diabetes is one of the most common and growing diseases in various countries that researchers around the world are trying to offer a way to prevent the disease. Diabetes is a chronic non-communicable disease. The growth of this disease is accompanied by an abnormal increase in glucose levels in the blood, which is divided into two categories include the first type of diabetes that occurs with insufficient production of insulin by the pancreas, and second is type 2 diabetes, which affects cells in effective response to produces insulin by losomedema. The disease has been closely followed up by the World Health Organization and the World Diabetes Federation since the rise of diabetes. By the end of 2015, nearly 392 million people will be diagnosed with diabetes. According to these figures, which are also growing daily according to reports from the World Diabetes Federation, its control, prevention and early detection can contribute to the development of this disease.

The world was drowning in big data before it even knew it existed. Big data had collected vast amounts of stored information at the time the term was raised which if properly analyzed, it could reveal valuable insights about the industry to which that particular data belongs. The convergence of artificial intelligence and big data already has far-reaching implications for how things are designed, create, and maintained.

Medical informatics owe the structures for processing, storing and disseminating information for various fields in medicine [1]. The main goal behind all efforts to classify, cluster, and extract features from existing data is to make the decision support system that helps people to identify and diagnose illnesses. Recently, data mining methods are widely used to identify and develop advanced diseases [2-4]. But recent research has shown a high-precision classifier system has not been provided to different data in the same conditions to date [5]. Data mining techniques are used as a prominent tool in medical databases. These methods can be used to index a new window to identify and diagnose diseases, which can contribute to the advancement of science and the reduction of disease in different societies. Nowadays classification is used as an appropriate tool in the field of medicine in order to discover patients' data and to provide predictable modeling. Classification is a tool used in decision making systems and to date, various classification methods have been created in various scientific fields for the purpose of discovering knowledge.

This article organized as follows: in section 2, we study some literature review of diabetes diagnosis based on data mining methods and describe their methods and also advantages and disadvantages. In section 3, we propose a new model to diagnose diabetes disease features and then determine classification method to evaluate our approach and compare it with recent models. In section 4, we represent a simulation based on MATLAB platform to show our proposed methods is more optimized than recent ones and at last, in section 5, we discuss and define a good conclusion to obtain and determine how our proposed method can solve diabetes diagnosis challenges.

2. Literature Review

Various studies have been conducted to diagnose diabetes with valid data. In [6], a nonlinear classification using fuzzy logic based on genetic algorithm provided on multiple data including diabetes data, blood pressure, breast cancer, and Iris data, which results indicate a relatively high accuracy of this study compared to other previous methods, including Naïve-Bayesian, regression, neural network, Radial Basis Function (RDF) and several other methods. The data used in the field of diabetes is the Indian PIMA Dataset. In [7], the use of the Levenberg Marquardt method was proposed to evaluate the effectiveness of reducing the error during the classification of diabetic data. This study used the PIMA Indian data set. The training algorithm is dynamically applied to the neural network to reduce the error and is given by continuous network training until the optimal stage is reached. The neural network used in this study is a Multi-Layered Perceptron (MLP) neural network and has been used to estimate the minimum error.

In [8], a comparison made for existing classifiers used to predict diabetes. A priori methods investigated such as decision tree, artificial neural network, logical regression and Naïve-Bayesian. The proposed method of Boogie and Boosting developed to improve the effectiveness and predictability of diabetes data. The data used in this study relates to the data of diabetic patients in Thailand. In [9], a method for predicting diabetes presented using the classification method in data mining which also performs an evaluation analysis. The use of the Adaboost method based on C4.5 decision tree used as a data trainer. The results indicated the proposed method is more efficient than the Boosting and Begging methods. The data used is also CPCSSN. In [10], presented the efficiency analysis for classification models for predicting diabetes. The results once with and once without the noise data noise was significant. Noise is the data that has not been normalized. Evaluation criteria include precision, sensitivity and data characteristics. The results indicated the randomized FURST method with more than 99% has a higher ability to classify the diabetes data set compared to the methods mentioned.

In [11], an automatic prediction system for diabetes mellitus presented based on a combination of Support Vector Machine (SVM) with wavelet transform based linear separation analysis. Actually, in the abbreviation, the name of the LDA-MWSVM method is named. This work consists of three steps: the stage of extracting the feature and the step of reducing the dimensions of the feature using the linear separation analysis method and the classification step using a combination of SVM model with wavelet transform. In the third step, the operation of predicting the correct results and performance of the proposed system is analyzed using sensitivity analysis, specificity, classification accuracy and confusion matrix. The classification accuracy of this system is close to 89.74%. The datasets used in this study are valid data from a website.

In [12], a fuzzy classification system based on an Ant Colony Optimization (ACO) algorithm proposed to predict diabetes. The purpose of this study is to use a classification system based on an ACO algorithm that can extract a set of fuzzy rules for predicting diabetes, termed FCS-ANTMINER. The accuracy of the classification method is estimated at 84.24%. The datasets used in this study are PIMA Indian data. In [13], a comparison has been made between two methods of classification of the MLP neural network and logical regression in the American people diabetes data set. The extraction operation is considered to be a critical operation in this study, since the correctness of this section can challenge the scientific comparison between the two methods of MLP neural network and logical regression. For this purpose, a constant algorithm has been used to extract the property, which is the Genetic Algorithm (GA). The results represented the sensitivity and specificity in the logic regression method are 0.9965 and 0.9946, and the sensitivity and specificity in the MLP neural network method is 0.9966 and 0.9918. It is clear that the MLP neural network has a better performance than the logical regression method for classifying diabetic data. The dataset used in this research is for the elderly in America.

In [14] early diagnosis of type II diabetes has been done using multiple classification systems to develop the accuracy of diagnosis for complex type II diabetes. A dynamic weighing schema, known as the weighted combination of multiple criteria presented in order to combine classification in decision making. This method does not include only local or global accuracy, but the participation in the classification and the generalized locational error for each classifier is also considered. The proposed method is called MFWC. The system was constructed using two types of diabetes data. In [15], a fuzzy classification method proposed using an Artificial Bee Colony (ABC) algorithm for diabetes. In this study, the mutation operator has been added from the ABC algorithm to improve performance, and if the best result obtained from the classification operation cannot be updated, the combined operator is added to improve the operation. In fact, the modified Honey Bee Colony (HBC) algorithm has been used in this study, which uses this algorithm as a new tool for the creation and optimization of the membership functions and rules derived from the data. The performance evaluation of the proposed method is based on the use of the classification rate, sensitivity, and specificity of the data, which is used for the 10-Fold-Cross-Vallidation method. The classification rate is 84.21%. The data used in this study is PIMA India.

In [16], diabetes prediction has been done using the Bacterial Nutrition Optimization (BNO) algorithm and the neural network. The data collection used in this study is PIMA India's diabetic data, which contains 768 data with eight characteristics which employs eight of the features. In [17], the use of the Convolutional Neural Network (CNN) and the Long-Short-Term-Memory (LSTM) for diagnosing diabetes according to the heart rate from the PIMA INDIA dataset have been estimated which results in the accuracy of the proposed method is estimated to be 93.6%. In [18], a powerful intelligent diagnostic system for diabetic disease presented from the PIMA INDIA dataset based on a hybrid algorithm called a fuzzy inference system based on adaptive logistics. Using the principles of extracting features and classifications has been considered in this research. The accuracy of the proposed method is relatively inadequate and is estimated at 88.03% and its computational complexity is high.

As review articles, it can be noted that [19] has compared the methods used to diagnose diabetes when using the PIMA INDIA dataset. The methods and algorithms that are considered include fuzzy logic, FCM, SVM, genetic algorithm, artificial neural network, and Principal Component Analysis (PCA) algorithm. Similarly, in [20], diagnosis of diabetes is based on PIMA INDIA data, which provides an overview of the methods, weaknesses and strengths along with their results. Some method studied such as SVM, Artificial Neural Network, Naïve-Bayesian, J48 Decision tree, Begging method, and combined method of genetic algorithm with SVM.

In [21], diabetes diagnosis and determine glucose level proposed based on two swarm intelligence optimization methods Moth Flame Optimization (MFO)-based Crow Search Algorithm (CSA) while using deep learning. The combination methods can count hidden neurons in multi-CNN layers to define minimum correlation between the features to avoid redundant information. Also, some fuzzy rules determined for define membership function which optimized based on MFO-CSA to classify diabetes features. Then Recurrent Neural Network (RNN) as deep learning technique used for predicting the range of enhanced data. RMSE and MASE used as evaluation criteria in clinical dataset. In [22], K-Nearest Neighbor (KNN) method used for diabetes detection based on Auto Encoder (AE) deep learning which tested on Pima India Diabetes Dataset. 5-fold cross-validation (FCV) used for results cross-validation and evaluation with the highest accuracy of 98.07%.

In [23], Extreme Gradient Boost (XGBoost) classifier methods used for diabetes diagnosis which applied in PPG signal dataset and gained 99.24% accuracy in Pima India Diabetes Dataset. Also, in [24] proposed a cluster-based and XGBoost classifier methods with 99.03% accuracy to detect diabetes in Pima India Diabetes Dataset. A combination method of SVM Hierarchical clustering and CNN deep learning used in [25] for diabetes detection with a good performance in terms of Receiver Operating characteristic (ROC). In [26], proposed deep Variational Auto-Encoder (VAE) model based Sparse Auto-Encoder (SAE) and CNN for feature augmentation in Pima India Diabetes Dataset with 92.31% accuracy. Among the issues that the reviewed articles refer to [6-26] for predicting, identifying, and categorizing diabetes, one can mention the following:

- Not using actual data and not specifying the type of data and number of available samples.
- The accuracy of the presented methods is relatively low.
- The number of evaluation methods to ensure the proposed method is low.
- There is no definite unit other than the precision of approaches for the proposed methods to be compared accordingly.
- The computational complexity of most methods is high.
- Most of the methods have no outputs in the paper and research, and only a series of numbers is presented.

3. Proposed Method

In this section, we combine three fuzzy-neural along with chaos theory and K-means tree. Clustering can be considered as the most important problem of uncontrolled learning. Therefore, as any other problem of this type, it is possible to find a structure in a set of unlabeled data. The K-means clustering, as well as the C-means method is known as segmentation method or data compression. The K-means cluster was invented in 1956. This method is based on the random selection of K from the initial cluster centers. These early cluster centers are updated as much as possible after selecting and data cycles. The central cluster can be randomly selected or can be based on previous information. Each point or data is attributed to a cluster. Finally, with respect to the central and primary cluster, it is recalculated and the convergence condition of this work is done. The K-means clustering set and the data vector are placed inside a number of predefined clusters, which is similar to Euclidean distance as a measure. The data vector in a cluster, the small Euclidean distance, is associated with one central vector, showing the midpoint of that cluster. The central vector is the data vector that belongs to the corresponding cluster. The process of the algorithm is as follows:

- Step 1: The algorithm starts with random initialization with C_i and selection of point c is done from all points.
- Step 2: Determine the membership matrix U in such a way that the u_{ij} elements are equal to one. Therefore, if the j is the data belonging to X_j, then the value is one. Otherwise it will be zero.
- Step 3: Calculate the cost function using equation (1). This relationship stops the algorithm if it is less than a threshold value.

$$J = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c}\left(\sum_{k,x_k \in C_i} \|X_k - C_i\|^2\right) \qquad (1)$$

- Step Four: Assign each point of data to the closest central cluster.
- Step 5: Update the main C_i main cluster by recalculating the central cluster as the average of all data points along with each cluster and defining a new U matrix. The parameters and options for the K-means algorithm include:

- ✓ Number of classes
- ✓ Initialization
- ✓ Measure distance
- ✓ Central cluster
- ✓ Check out and meet the condition of termination

Despite the fact that the K-means algorithm is terminated, the final solution is not the same and is not always the optimal answer. The weakness of the K-means method can be explained by the fact that the number of clusters is constant, since once the selection k remains, the cluster centers K remain. Of course, this can be solved by removing and removing waste clusters. Any sample that is not sampled at the center of the sample cluster can be removed, and the need for a new central cluster can be formed. The problem of choosing the number of clusters remains unaltered, but it can be counted greatly by using the selection of a large enough k. Finally, for analyzing large data sets, this method is not sufficiently used, since at each step of this method it is necessary to calculate the distance between each pair of data and even calculating all distances seems compulsory. Therefore, since this method is a chaotic algorithm, then it depends on the initial conditions, which leads to the convergent and optimal algorithm.

The chaotic mode begins with a population of randomly generated random solutions. A single solution is displayed through a simple string of butterfly effect. The quality of each butterfly effect is evaluated using the fitness function based on initial condition. The valid butterfly effect is a binary string of length K, for example the number of potential positions. The butterfly effect operations always lead to the production of binary strings of length K. Hence, it is not necessary to check the validity of butterfly effect in each replication.

The proposed butterfly effect encoding system is a randomly generated binary string length K. Therefore, encoding a butterfly effect for $O(K)$ and creating an initial population with P_{size} size takes $O(P_{size} \times K)$. The fitness value of a butterfly effect can be calculated over time $O(N \times K)$, where N is the number of target points. One-point intermixing operation has also been used, whose complexity to produce two chromosomes of the child is equal to $O(K)$. In the mutation process, the position of one of the genes is randomly selected and its value changes to zero. Hence, the self-similarity operation apply in a constant time, such as $O(1)$. It should be noted that the processes of fusion and uniqueness repeat until the end criteria are fulfilled, and after each repetition, the value of the newly generated butterfly effect is calculated and at this stage it must be determined which parent or child butterfly effect will be transferred to the next generation. Therefore, the total process of repeating the mixing and mutation will take place at time $O(1 \times (N \times K + K))$.

Then the fuzzy-neural enters the work to combine with chaotic K-means tree. The purpose of computing information is to use good candidate solutions to influence the design of new candidate solutions by positioning problem optimizations. This is done by constructing candidate solutions randomly in step-by-step levels, each component is randomly drawn from the quality solution memory and configure from a high-quality solution memory or randomly Problem areas are assigned. Candidate solution memory is random at the start of the work, and the criteria for adopting an intelligence method to prove new candidate solutions are only used when they have developed the target's status by replacing the existing member.

Fuzzy-neural or ANFIS learning has a function similar to neural networks. Enhanced learning methods provide a way to create a fuzzy modeling procedure to learn information from a dataset. Fuzzy logic calculates the parameters of the membership function so that the fuzzy inference system matches the input and output data sets. It is in this direction that ANFIS is used. The way of relationship between fuzzy logic and neural network has led to the creation of different types of systems. Many believe that applying the word ANFIS to all these combinations is not correct; because some of these combinations have a complementary relationship with each other and instead of each of these components, other systems such as decision tree, evolutionary algorithm and so on can be substituted. In other words, ANFIS is a combination of a neural network and a fuzzy inference system, in which the neural network is used to determine the parameters of the fuzzy system. The purpose of determining fuzzy system parameters by neural network is automatic determination of fuzzy parameters such as fuzzy rules or membership functions of fuzzy sets. In contrast to ANFIS, there is a fuzzy neural network in which fuzzy logic is used to improve the performance of the neural network. In this network, fuzzy logic is a branch and is used only to improve the conditions of the neural network or to add the concept of uncertainty to the network.

Due to the fact that GENEFIS3 based on ANFIS was used in the command line in this research, the advantages of GENEFIS3 are considered important. Among these advantages in fuzzy clustering is that the data can belong to more than one cluster with different degrees of membership functions. GENEFIS3 performs its tasks by extracting a series of rules that model data behaviors. The rule extraction method first uses the FCM function, which is used to determine the fuzzy set numbers and membership functions for all previous sections. The number of clusters is collected by comprehensive search method. This prediction is only a weighted sum between the last observation x_t and the prediction of the penultimate period F_t . In this method, the next period's demand is estimated by using the equation (2), where $0 < \alpha < 1$ is called the smoothing constant.

$$F_{t+1} = \alpha x_t + (1-\alpha)F_t \tag{2}$$

Due to the existence of these recursive relations between F_t and F_{t+1}, it is possible to display F_{t+1} in another way similar to equation (3). It is clear that in this form of expressing the relationship, exponential smoothing assigns the most weight to x_t and lower weights to previous observations. In addition, this relationship will be a simple method in estimating the demand of the next period because it does not need to keep the data before period t.

$$F_{t+1} = \alpha x_t + \alpha(1-\alpha)x_{t-1} + \alpha(1-\alpha)^2 x_{t-2} + \cdots \tag{3}$$

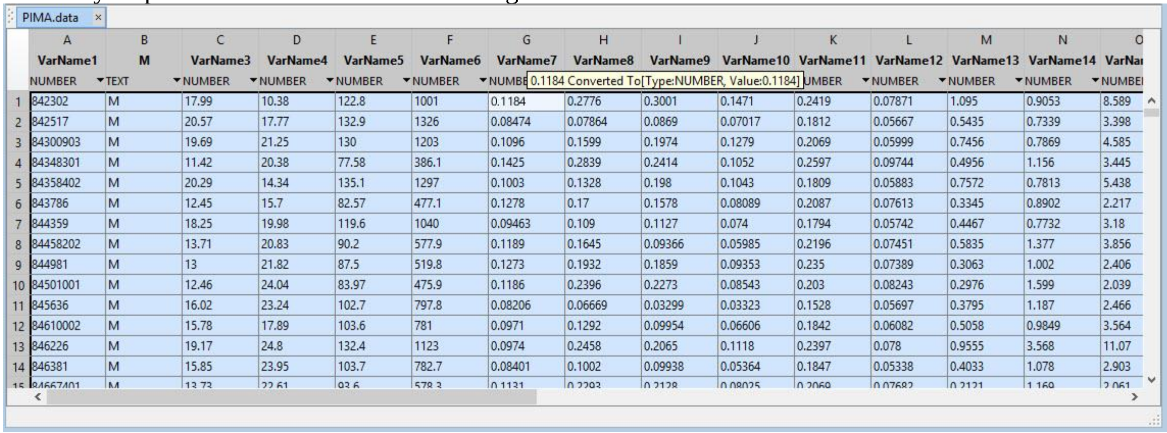
All that is required is x_t and the prior prediction of F_t . The relationship of exponential smoothing can be expressed in another way similar to the equation (4).

$$F_{t+1} = F_t + \alpha(x_t - F_t) \tag{4}$$

This equation shows the prediction for period $t+1$ is equal to the sum of the prediction of the penultimate period t and the product of the prediction error in period t by a discount factor α.

4. Simulation and Results

The simulation will be done in the MATLAB environment. The data from this research is the use of PIMA INDIAN with 768 data. The parameters that are included in this dataset as diagnostic factors for diabetes include the number of pregnant people, the level of blood sugar, systolic, scaling, insulin, age, body mass index, and inheritance factors in a family. It should be noted that three features including blood glucose levels, insulin and systolic intake, have been identified as the three main attributes in identifying and diagnosing diabetes in this study. A portion of this data is shown in Fig. 1.



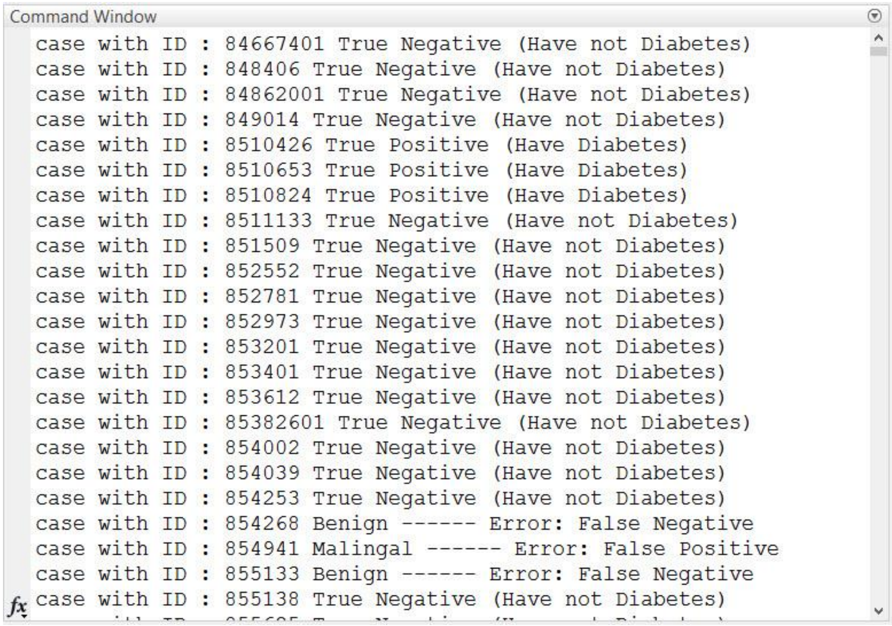
PIMA.data

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
	VarName1	M	VarName3	VarName4	VarName5	VarName6	VarName7	VarName8	VarName9	VarName10	VarName11	VarName12	VarName13	VarName14	VarNa
	NUMBER	TEXT	NUMBER	NUMBER	NUMBER	NUMBER	NUMBE 0.1184 Converted To[Type:NUMBER, Value:0.1184]				UMBER	NUMBER	NUMBER	NUMBER	NUMBE
1	842302	M	17.99	10.38	122.8	1001	0.1184	0.2776	0.3001	0.1471	0.2419	0.07871	1.095	0.9053	8.589
2	842517	M	20.57	17.77	132.9	1326	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667	0.5435	0.7339	3.398
3	84300903	M	19.69	21.25	130	1203	0.1096	0.1599	0.1974	0.1279	0.2069	0.05999	0.7456	0.7869	4.585
4	84348301	M	11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414	0.1052	0.2597	0.09744	0.4956	1.156	3.445
5	84358402	M	20.29	14.34	135.1	1297	0.1003	0.1328	0.198	0.1043	0.1809	0.05883	0.7572	0.7813	5.438
6	843786	M	12.45	15.7	82.57	477.1	0.1278	0.17	0.1578	0.08089	0.2087	0.07613	0.3345	0.8902	2.217
7	844359	M	18.25	19.98	119.6	1040	0.09463	0.109	0.1127	0.074	0.1794	0.05742	0.4467	0.7732	3.18
8	84458202	M	13.71	20.83	90.2	577.9	0.1189	0.1645	0.09366	0.05985	0.2196	0.07451	0.5835	1.377	3.856
9	844981	M	13	21.82	87.5	519.8	0.1273	0.1932	0.1859	0.09353	0.235	0.07389	0.3063	1.002	2.406
10	84501001	M	12.46	24.04	83.97	475.9	0.1186	0.2396	0.2273	0.08543	0.203	0.08243	0.2976	1.599	2.039
11	845636	M	16.02	23.24	102.7	797.8	0.08206	0.06669	0.03299	0.03323	0.1528	0.05697	0.3795	1.187	2.466
12	84610002	M	15.78	17.89	103.6	781	0.0971	0.1292	0.09954	0.06606	0.1842	0.06082	0.5058	0.9849	3.564
13	846226	M	19.17	24.8	132.4	1123	0.0974	0.2458	0.2065	0.1118	0.2397	0.078	0.9555	3.568	11.07
14	846381	M	15.85	23.95	103.7	782.7	0.08401	0.1002	0.09938	0.05364	0.1847	0.05338	0.4033	1.078	2.903
15	84667401	M	13.73	22.61	93.6	578.3	0.1131	0.2293	0.2128	0.08025	0.2069	0.07682	0.2121	1.169	2.061

Fig. 1. PIMA INDIA Dataset

The first column is the identifier of each patient, in the second column, M is the Malignant Representative or Malingal, and B is Benign or benign agent. Other cases in the columns are other features of diabetes based on these features and its intervals are given by classification and extraction of attributes that can properly diagnose diabetes. These values are not handled manually because the goal is not numerical data, but the use of references in this field is essential that is used from reference [18-14].

According to the proposed method of this study, which uses a combination of fuzzy-neural chaotic K-means tree algorithm, diagnosis of diabetes is to be done on the data. For this purpose, after applying the proposed approach, it is necessary to show the results of the classification of the proposed method. To this end, every user who has taken a test from him has an identifier in the PIMA INDIA dataset. Based on these identifiers, classification classes are created that include diabetes, non-diabetes, and suspicious. In some areas, due to the main characteristic of this fat-based study, if the amount of fat is over and in the range, the malignant state of diabetes is benign and, if it is appropriate, diabetes, it is considered benign. Figure (2) shows the result of this classification.



```
case with ID : 84667401 True Negative (Have not Diabetes)
case with ID : 848406 True Negative (Have not Diabetes)
case with ID : 84862001 True Negative (Have not Diabetes)
case with ID : 849014 True Negative (Have not Diabetes)
case with ID : 8510426 True Positive (Have Diabetes)
case with ID : 8510653 True Positive (Have Diabetes)
case with ID : 8510824 True Positive (Have Diabetes)
case with ID : 8511133 True Negative (Have not Diabetes)
case with ID : 851509 True Negative (Have not Diabetes)
case with ID : 852552 True Negative (Have not Diabetes)
case with ID : 852781 True Negative (Have not Diabetes)
case with ID : 852973 True Negative (Have not Diabetes)
case with ID : 853201 True Negative (Have not Diabetes)
case with ID : 853401 True Negative (Have not Diabetes)
case with ID : 853612 True Negative (Have not Diabetes)
case with ID : 85382601 True Negative (Have not Diabetes)
case with ID : 854002 True Negative (Have not Diabetes)
case with ID : 854039 True Negative (Have not Diabetes)
case with ID : 854253 True Negative (Have not Diabetes)
case with ID : 854268 Benign ------ Error: False Negative
case with ID : 854941 Malingal ------ Error: False Positive
case with ID : 855133 Benign ------ Error: False Negative
case with ID : 855138 True Negative (Have not Diabetes)
```

Fig. 2. The result of the classification with the aim of diagnosing diabetes with the proposed approach

According to Figure 2, in the first line on the command line, it is shown that the person with the identifier 84667401 does not have diabetes, so according to the accuracy criterion, TN will be. In the fifth line, the person with ID 8510426 has diabetes, so according to the accuracy criterion, TP will be. Also, based on this output, from the bottom of the second case, the user is shown with the ID 855133, which is the user of FN, that is, benign type diabetes. Also, from the bottom of the third, the user is shown with the identifier 854941, which is the user of FP, that is, malignant diabetes. After combining the classification and extraction function based on the combined use of the genetic algorithm, the K-means method and the Harmonic Search algorithm, the results of the evaluation criteria are shown. Table (1) shows the values obtained for each evaluation method during the combined operation of classification and feature extraction.

Table (1), the results obtained from the proposed method

Accuracy (%)	94.87%
Specificity (%)	93.97%
Sensitivity (%)	86.39%

The proposed method will be compared in a number of ways, which will be based on evaluation criteria, including sensitivity, rate of attributes and rate of classification or accuracy, and are shown in Fig. 3. It should be noted that the reference articles are reference articles [18] and [17].

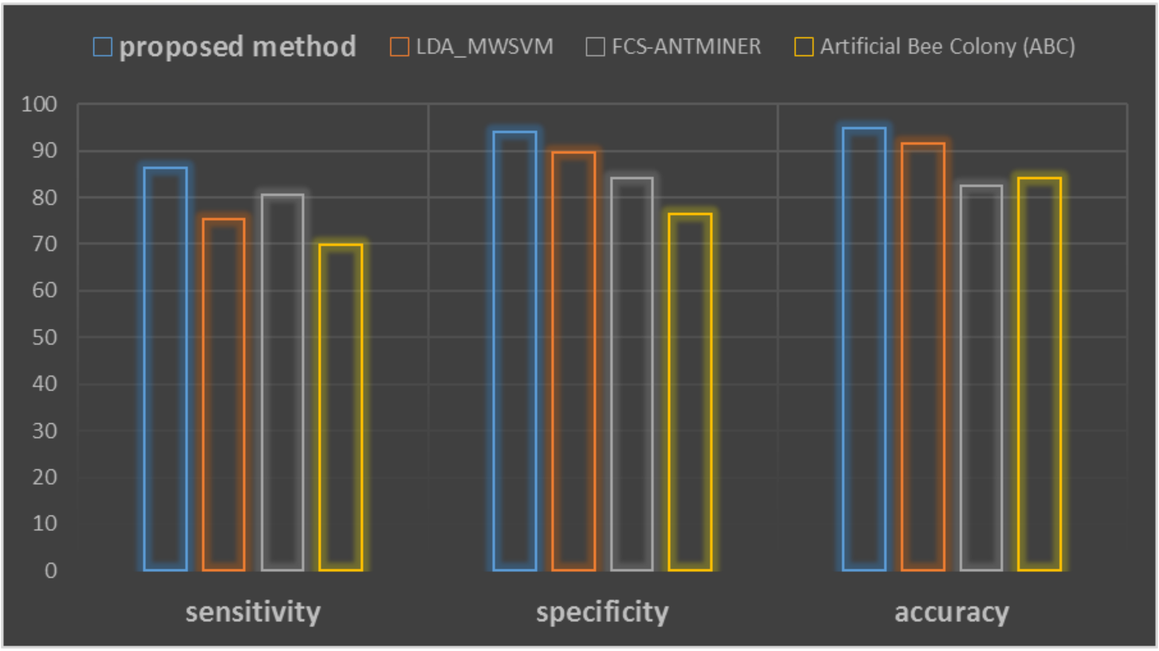


Fig. 3 shows the results compared with the other three methods

It is clear that our proposed method in this study has a higher ability to classify and therefore predict diabetes. We consider three evaluation criteria as comparisons that include the sensitivity, the amount of attributes, and the rate of classification or accuracy that most of the previous methods use to compare the three methods. It is worth noting that under equal conditions, the direct use of other methods from a dataset with the same number of attributes or less, a better comparison can be made. It is worth noting that 768 data values from the PIMA India dataset with its three selected features were considered in this study, the classification time was 24.44 seconds. But a comparison has been made by adding two newer approaches, which can be seen in Table 2 based on the accuracy criterion in percent.

Table (2), comparison in terms of accuracy with previous methods

Methods and References	Accuracy (%)
LDA-MWSVM [14]	91.547 %
FCS-ANTIMINER [15]	82.412 %
ABC [16]	84.21 %
CNN-LTSM [17]	93.60 %
LANFIS [18]	88.03 %
K-Nearest Neighbor (KNN) and Auto Encoder (AE) deep learning [22]	98.07 %
Extreme Gradient Boost (XGBoost) classifier [23]	99.24 %
Cluster-based and XGBoost classifier [24]	99.03 %
Deep Variational Auto-Encoder (VAE) model based Sparse Auto-Encoder (SAE) and CNN [26]	92.31 %
Proposed Method	94.87 %

5. Conclusion

Diabetes mellitus as a disorder of all ages threatens people's health. Recognizing its early symptoms that occurs as a cover can prevent dangers such as discontinuation of the body, blindness, and so on. Hence, providing systems that can help doctors to diagnose and diagnose diabetes early is important. In this study, a new method was used to classify diabetic data that can be used to identify, diagnose and predict diabetes. The data used in this work was a PIMA India dataset that we use of its three features. Combination method used in this article for clustering diabetes data based on fuzzy-neural chaotic K-means tree to extract and select features and classification. The results compared with the previous methods in terms of accuracy evaluation criteria, and the results indicated the proposed method is superior to the previous methods. The accuracy of the proposed method is 94.87%, which has a relative advantage over the other methods presented in the articles under the same conditions of data set usage.

References

[1] Shortliffe, E. H., 1990. Medical Informatics: Computer Applications in Medicine, Addison-Wesley.
[2] Botstein, D. and Risch, N., 2003. Discovering genotypes underlying human phenotypes: past successes for Mendelian disease, Future approaches for complex Disease, Nat. Genet., 228-237.
[3] Huang, Y., McCullagh, P., Black, N., and Harper, R., 2005. Feature selection and classification model construction on type 2 diabetic patient’s data, Adv. Data Min., 153-162.
[4] Tama, B., and Rodiyatul F., 2010. An early detection method of type-2 diabetes mellitus in public hospital, Telkomnika, 287-294.
[5] Wolpert, D. H., 1996. The lack of a priori distinctions between learning algorithms, Neural Computing, 1341-1390.
[6] Fang, Hua, L.Rizzo, Maria, Wang, Honggang, Andrews Espy, Kimberly, and Wang, Zhenyuan, 2010. A new nonlinear classifier with a penalized signed fuzzy measure using effective genetic algorithm, Pattern Recognition, 43, 1393-1401.
[7] Khan, Nawaz, Gaurav, Dhara, and Kandl, Thomas, 2013. Performance Evaluation of Levenberg-Marquardt Technique in Error Reduction for Diabetes Condition Classification, International Conference on Computational Science (ICCS), Procedia Computer Science, 18, 2629-2637.
[8] Nai-arun, Nongyao, and Moungmai, Rungruttikarn, 2015. Comparison of Classifiers for the Risk of Diabetes Prediction, 7th International Conference on Advances in Information Technology, Procedia Computer Science, 69, 132-142.
[9] Perveen, Sajida, Shahbaz, Muhammad, Guergachi, Aziz, and Keshavjee, Karim, 2106. Performance Analysis of Data Mining Classification Techniques to Predict Diabetes, Symposium on Data Mining Applications (SDMA), Riyadh, Saudi Arabia, Procedia Computer Science, 82, 115-121.
[10] Kandhasamy, J. Pradeep, and Balamurali, S., 2015. Performance Analysis of Classifier Models to Predict Diabetes Mellitus, Procedia Computer Science, 47, 45-51.
[11] Calisir, Duygu, and Dogantekin, Esin, 2011. An automatic diabetes diagnosis system based on LDA-Wavelet Support Vector Machine Classifier, Expert Systems with Applications, 48, 8311-8315.
[12] Fathi Ganji, Mostafa, and Saniee Abadeh, Mohammad, 2011. A fuzzy classification system based on Ant Colony Optimization for diabetes disease diagnosis, Expert Systems with Applications, 38, 14650-14659.
[13] Upadhyaya, S., Farahmand, K., and Baker-Demaray, T., 2013. Comparison of NN and LR classifiers in the context of screening Native American elders with diabetes, Expert Systems with Applications, 40, 5830-5838, 2013.
[14] Zhu, Jia, Xie, Qing, and Zheng, Kai, 2015. An improved early detection method of type-2 diabetes mellitus using multiple classifier system, Information Sciences, 292, 1-14.
[15] Beloufa, Fayssal, and Chikh M. A., 2013. Design of fuzzy classifier for diabetes disease using Modified Artificial Bee Colony algorithm, Computer Methods and Program in Biomedicine, 112, 92-103.
[16] Rabina, and Sharma, Meenakshi, 2016. Diabetes Prediction by using Bacterial Foraging Optimization Algorithm and Artificial Neural Network, International Journal of Computer Science and Information Technology & Security (IJCSITS), 6, 1.
[17] Swapna, G., Soman, Kp, and Vinayakumar, R., 2018. Automated detection of diabetes using CNN and CNN-LSTM network and heart rate signals, Procedia Computer Science, 132, 1253-1262.
[18] Ramezani, Rohollah, Maadi, Mansoureh, and Khatami, Seyedeh Malihe, 2017. A novel hybrid intelligent system with missing value imputation for diabetes diagnosis, Alexandria Engineering Journal, in press, corrected proof, Available online 19 April 2017.
[19] Gujral, Sakshi, 2017. Early Diabetes Detection using Machine Learning: A Review, IJIRST–International Journal for Innovative Research in Science &Technology, 3, 10.
[20] Fatima, Meherwar, and Pasha, Maruf, 2017. Survey of Machine Learning Algorithms for Disease Diagnostic, Journal of Intelligent Learning Systems and Applications, 9, 1-16.
[21] Naveena, Somasundaram, and Bharathi, Ayyasamy, 2022. A new design of diabetes detection and glucose level prediction using moth flame-based crow search deep learning, Biomedical Signal Processing and Control, 77.
[22] Suyanto, Suyanto, Meliana, Selly, Wayhuninrum, Tenia, and Khomsah, Siti, 2022. A new nearest neighbor-based framework for diabetes detection, Expert Systems with Applications, 199.
[23] Prabha, Anju, Yadav, Jyoti, Rani, Asha, and Singh, Vijander, 2021. Design of intelligent diabetes mellitus detection system using hybrid feature selection based XGBoost classifier, Computers in Biology and Medicine, 136.
[24] Hassan, Mehedi, Mollick, Swarnali, and Yasmin, Farhana, 2022. An unsupervised cluster-based feature grouping model for early diabetes detection, Healthcare Analytics, 2.
[25] Fang, Jinhai, Xie, Zuoling, Cheng, Haitao, Fan, Bin, Xu, He, and Li, Pen, 2022. Anomaly detection of diabetes data based on hierarchical clustering and CNN, Procedia Computer Science, 199, 71-78.
[26] Teresa Garcia Ordas, Maria, Nbenavides, Carmen, Alberto Benitez-Andrades, Jose, Alaiz-Moreton, Hector, and Garcia-Rodriguez, Isaias, 2021. Diabetes detection using deep learning techniques with oversampling and feature augmentation,Computer Methods and Programs in Biomedicine, 202.