Transactions on Machine Intelligence 6(2) (2023) 104-113

ISSN Online: 2821-1693

Transactions on Machine Intelligence

Journal Homepage: https://tmachineintelligence.ir/index.php/TMI

Diabetes Diagnosis from Big Data using Fuzzy-Neural Chaotic Tree

B. Saleh[1,*], H. Hasanpour[2]

[1] Assistant Professor, Department of Information Technology, Sabzevar Branch, Islamic Azad University, Sabzevar, Iran
[2] Department of Computer Engineering, Sabzevar Branch, Islamic Azad University, Sabzevar, Iran

ARTICLE INFO

Article History:
Received 10 February 2023
Received in revised form 14 March 2023
Accepted 5 May 2023
Available online 11 June 2023

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

ABSTRACT

Today, diabetes is a recognized global health concern. Global statistics show an increasing prevalence of the disease, posing a challenging issue for modern medicine. In response, computer science has proposed various methods to diagnose and predict diabetes. Nonetheless, researchers continue to work on resolving outstanding issues and errors. Data mining is used as a technical method for identifying and extracting new knowledge from data. This study introduces a novel approach for categorizing diabetic data that consists of three stages. Firstly, pre-processing is conducted, where data normalization procedures are applied. Subsequently, attribute extraction and selection are carried out. Finally, data mining principles are utilized for classification. The classification results obtained can be utilized to predict diabetes in various individuals. Evaluation of the results involves adherence to certain standards such as sensitivity, specificity, and accuracy. Our recommended approach, which combines chaotic fuzzy-neural with K-means tree, proves more effective than previous techniques, as confirmed by the results.

1. INTRODUCTION

Diabetes is a prevalent and expanding health issue in numerous countries, prompting global researchers to develop preventive measures for this chronic non-communicable disease. The disease's progression leads to an anomalous spike in blood glucose levels, categorizing it into two types: the first, arising from a decrease in insulin production by the pancreas, and the second, eliciting an ineffective cellular response to insulin production by losomedema. Diabetes has been under the strict surveillance of both the World Health Organization and the World Diabetes Federation since its emergence. By the end of 2015, approximately 392 million individuals will receive a diagnosis of diabetes. These numbers are consistently increasing, as reported by the World Diabetes Federation. Control, prevention, and early detection of the disease can aid in its progress.

The emergence of big data overwhelmed society before its recognition. Big Data collected a significant amount of stored information at the time of its inception. If analyzed properly, this information could provide valuable insights about the industry to which the data belongs. The convergence of Artificial Intelligence and Big Data has far-reaching implications for design, creation, and maintenance processes.

* Corresponding Author: banafshehsaleh07@gmail.com
Assistant Professor, Department of Information Technology, Sabzevar Branch, Islamic Azad University, Sabzevar, Iran

http://dx.doi.org/10.47176/TMI.2023.104

Medical informatics owe the structures for processing, storing, and disseminating information for various fields in medicine [1]. The primary objective of all efforts to classify, cluster, and extract features from existing data is to create a decision support system that aids in identifying and diagnosing illnesses. Currently, data mining methods are extensively employed to identify and develop advanced diseases [2-4]. Recent research has shown that a high-precision classification system has not yet been developed for different data under similar conditions [5]. Data mining techniques serve as a primary tool in medical databases, which can index new windows to identify and diagnose diseases that contribute to the advancement of science and the reduction of disease in various societies. Nowadays, classification serves as an essential tool in medicine to extract patient data and create predictable models. It aids decision-making systems, and a range of classification methods have been developed in various scientific fields to explore novel findings.

This article is organized as follows: in Section 2, we conduct a literature review of diabetes diagnosis utilizing data mining methods. We describe the methods, advantages, and disadvantages of these studies. In Section 3, we introduce our new model for diagnosing diabetes disease features and subsequently determine a classification method to evaluate our approach while comparing it with recent models. In Section 4, a simulation based on the MATLAB platform depicts the optimization of our proposed methods in comparison to recent ones. Finally, in Section 5, we analyze and define a suitable conclusion to ascertain how our proposed method can address challenges related to diabetes diagnosis.

2. LITERATURE REVIEW

Various studies have conducted valid diagnoses of diabetes. In [6], a nonlinear classification utilizing fuzzy logic based on a genetic algorithm provided results on multiple data sets that include diabetes, blood pressure, breast cancer, and Iris data. The results indicate relatively higher accuracy compared to other previous methods such as Naïve-Bayesian, regression, neural network, Radial Basis Function (RDF), and several others. The dataset utilized in diabetes research is the PIMA Indian dataset. Reference [7] proposed the use of the Levenberg Marquardt method to assess its effectiveness in reducing error during diabetic data classification. This study utilized the same dataset, and the training algorithm is dynamically applied to the neural network to minimize error. The network is continuously trained until reaching the optimal state. The Multi-Layered Perceptron (MLP) neural network was utilized in this study to estimate the minimum error.

A comparison between existing classifiers for predicting diabetes was conducted in [8]. Several a priori methods were investigated, including decision trees, artificial neural networks, logical regression, and Naïve-Bayesian. The proposed Boogie and Boosting method were developed to enhance the effectiveness and predictability of diabetes data. The study data pertains to diabetic patients in Thailand. In [9], the classification method in data mining was employed for predicting diabetes by means of an evaluation analysis. The Adaboost method was utilized as the data trainer, based on C4.5 decision tree. The obtained results depicted the proposed approach as displaying greater efficiency compared to Boosting and Begging techniques.

The data utilized is also from CPCSSN. An efficiency analysis for classification models to predict diabetes was presented in [10]. The comparison of results with and without noise data was significant. Noise data refers to the data that has not been normalized. Precision, sensitivity, and data characteristics were included in the evaluation criteria. The randomized FURST method yielded over 99% accuracy and had a higher ability to classify the diabetes dataset when compared to other methods mentioned.

In [11], an automatic prediction system for diabetes mellitus was presented using a combination of Support Vector Machine (SVM) and wavelet transform-based linear separation analysis. The LDA-MWSVM method is the abbreviation's name. The language used is clear, objective, and free of emotional or ornamental language. The text adheres to formal register, uses precise word choice, and follows conventional and logical structure. The grammar, spelling, and punctuation are correct. This study comprises three main steps: first, the extraction of feature; second, dimension reduction of features using the linear separation analysis technique; and third, classification using a combination of SVM model alongside wavelet transform. The last step involves the analysis of the operation's efficacy by assessing the proposed system's sensitivity, specificity, classification accuracy, and confusion matrix.

The system's classification accuracy approaches 89.74%. The datasets employed in this study are valid data retrieved from a website.

In [12], a study proposed a fuzzy classification system that is based on the Ant Colony Optimization algorithm, and this system was aimed at predicting diabetes. The research objective is to use the ACO-based classification system and extract a set of fuzzy rules for predicting diabetes called FCS-ANTMINER. The study's accuracy for the classification method is estimated at 84.24%. PIMA Indian data sets were employed in the study. In the study outlined in [13], the accuracy of the extraction operation in the American diabetes dataset was found to be critical to a scientific comparison of two classification methods: the MLP neural network and logical regression. A Genetic Algorithm (GA) was utilized to extract the property in question. The obtained results indicate that the MLP neural network method is superior to the logic regression method in terms of sensitivity and specificity, achieving 0.9966 and 0.9918, respectively. In contrast, the logic regression method achieved 0.9965 sensitivity and 0.9946 specificity. These findings demonstrate the MLP neural network's better performance in classifying diabetic data. The data set utilized in this study involves elderly individuals residing in the United States.

Earlier diagnoses of type II diabetes have been executed utilizing various classification systems to enhance the accuracy of identifying complex type II diabetes, as shown in [14]. To consolidate the decision-making process, a dynamic weighing schema referred to as the weighted combination of multiple criteria was introduced. This approach considers not only local or global accuracy, but also takes into account the involvement in classification and the generalized locational error for every classifier.

The MFWC method was utilized to construct a system based on two distinct types of diabetes data. In [15], a fuzzy classification approach employing the Artificial Bee Colony (ABC) algorithm was proposed for diabetes. This study introduces a mutation operator from the ABC algorithm to enhance the classification method's efficacy. If the best result from the classification operation cannot be improved, a combined operator is utilized. The modified Honey Bee Colony (HBC) algorithm was employed in this study as a novel tool for the creation and optimization of membership functions and rules derived from data.

The proposed method's performance was evaluated using the 10-Fold-Cross-Validation method, which utilized the classification rate, sensitivity, and specificity of the data. The resulting classification rate was 84.21%. The data utilized in this study came from PIMA India.

In reference [16], researchers utilized the Bacterial Nutrition Optimization (BNO) algorithm and neural network to predict diabetes. Data from PIMA India's diabetic database, consisting of 768 records and eight characteristic variables, was utilized in this study. In study [17], researchers estimated the accuracy of a proposed method for diagnosing diabetes using the Convolutional Neural Network (CNN) and Long-Short-Term-Memory (LSTM) based on heart rate data from the PIMA INDIA dataset, resulting in an estimated accuracy of 93.6%. Study [18] presented a powerful intelligent diagnostic system for diabetic disease, using a hybrid algorithm called a fuzzy inference system based on adaptive logistics, also utilizing the PIMA INDIA dataset. The research has applied feature extraction and classification principles. According to estimations, the proposed method's accuracy is relatively insufficient, being calculated at 88.03%, and it has a high computational complexity [19].

Review articles have compared the methods employed to diagnose diabetes using the PIMA INDIA dataset. These methods include fuzzy logic, FCM, SVM, genetic algorithm, artificial neural network, and Principal Component Analysis (PCA) algorithm. Additionally, the diagnosis of diabetes in [20] is based on PIMA INDIA data, providing an overview of the techniques' weaknesses, strengths, and outcomes. Several methods were studied for diabetes diagnosis and determining glucose levels such as SVM, Artificial Neural Network, Naïve-Bayesian, J48 Decision tree, Begging method, and the combined genetic algorithm with SVM.

In reference [21], proposed optimization methods for swarm intelligence Moth Flame Optimization (MFO)-based Crow Search Algorithm (CSA) were utilized in conjunction with deep learning.

The method combines the counting of hidden neurons in multi-CNN layers to determine a minimum correlation between features, thus preventing redundant information. Additionally, fuzzy rules were established to define optimized membership functions based on MFO-CSA for classifying diabetes features. A Recurrent Neural Network (RNN) was employed as a deep learning technique to predict the range of enhanced data. RMSE and

MASE served as evaluation criteria in a clinical dataset. In [22], the Auto Encoder (AE) deep learning methodology was utilized to test the K-Nearest Neighbor (KNN) approach for diabetes detection on the Pima Indian Diabetes Dataset. Results were obtained using 5-fold cross-validation (FCV) for 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.

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

- 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 applies 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 network is integrated with the chaotic K-means algorithm. The objective of information computation is to influence the design of new candidate solutions utilizing good quality solutions by identifying problem optimizations. To achieve this, the candidate solutions are generated randomly in a step-by-step manner, with each component drawn randomly from the high-quality solution memory and configured with random assignment of problematic areas. At the outset of the work, the candidate solution memory is random, and intelligence methods are only adopted based on objective criteria. These methods are employed to validate new candidate solutions only after they have replaced the existing member and developed the target's status.

Fuzzy-neural or ANFIS learning functions much like neural networks, while improved learning methods allow a fuzzy modeling procedure that can glean information from a dataset. Fuzzy logic computes the membership function's parameters for the purpose of matching the input and output datasets of the fuzzy inference system. ANFIS is utilized in this regard. The connection between fuzzy logic and neural networks has resulted in the development of various system types. This is due to some of these combinations having a complementary relationship with each other and other systems such as decision tree, evolutionary algorithms, etc., can function in place of each of these components. Many experts argue that using the term ANFIS for all the aforementioned combinations is inaccurate. In simpler terms, ANFIS is a blend of a neural network and a fuzzy inference system that uses the former to establish the parameters of the latter.

The goal of utilizing neural networks to determine fuzzy system parameters is to automatically identify fuzzy parameters, including fuzzy rules or membership functions of fuzzy sets. Contrary to ANFIS, a fuzzy neural network incorporates fuzzy logic to enhance the overall functionality of the neural network. This type of network involves fuzzy logic as a subsidiary branch that is utilized solely to augment the network's conditions or to introduce the concept of uncertainty.

Since we used GENEFIS3, which is based on ANFIS in this research, we consider GENEFIS3's advantages to be significant. One of these benefits in fuzzy clustering is that data can belong to more than one cluster with varying levels 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	NUMBEI
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 lists the patient's identifier. The second column indicates whether the case is Malignant or Benign. Other columns show additional features of diabetes. These features are classified and extracted to

effectively diagnose diabetes. These values are not manually processed because the goal is not to obtain numerical data. However, references are essential in this field and are used from reference [18-14].

The proposed method of this study utilizes a combination of fuzzy-neural chaotic K-means tree algorithm to diagnose diabetes in the data. As such, it is necessary to present the results of the classification of the proposed method after applying this approach. To achieve objectivity, the PIMA INDIA dataset records an identifying code for every user, which enables the creation of classification categories for diabetes, non-diabetes, and suspicious cases. In certain areas of the study, if the quantity of fat is within the appropriate range, diabetes is classified as benign and non-diabetes as malignant. Figure (2) displays the resultant classification.

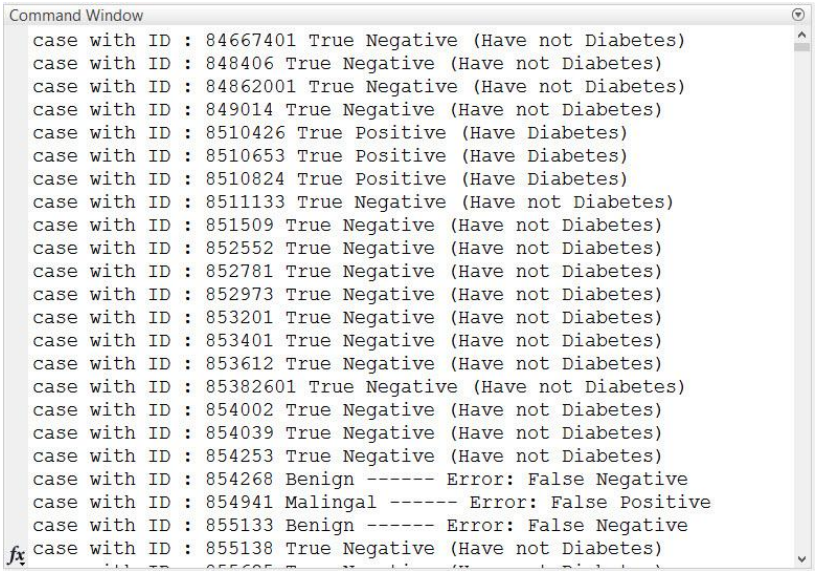

Fig. 2. The result of the classification with the aim of diagnosing diabetes with the proposed approach

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].

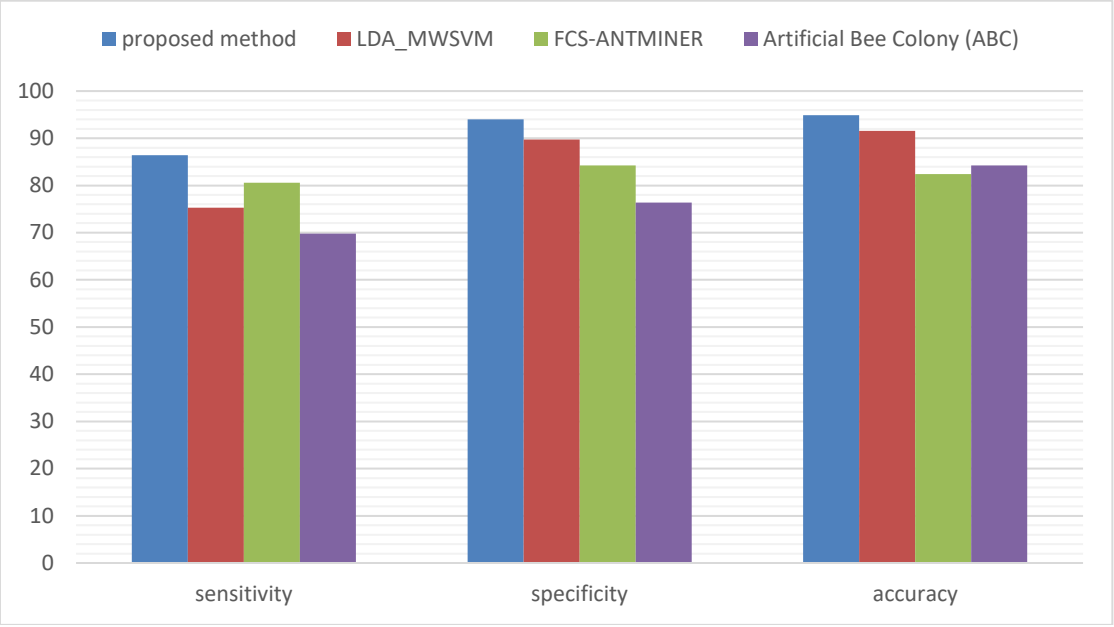

Fig. 3. Shows the results compared with the other three methods

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 number 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 is a health disorder that poses a threat to people of all ages. Early recognition of its symptoms that manifest covertly can prevent severe complications such as organ failure and vision loss. Therefore, development of diagnostic systems that enable early detection of diabetes is crucial. This study introduces a new method for classification of diabetic data that can aid in identification, diagnosis and prediction of diabetes. The dataset utilized for this research was the PIMA India dataset, focusing on three specific features. The article employs a combination method to cluster diabetes data, utilizing a fuzzy-neural chaotic K-means tree to extract and select features for classification. The results were then compared to previous methods using accuracy

evaluation criteria, indicating that the proposed method outperforms the previous approaches. The proposed method's accuracy is 94.87%, providing a relative advantage over other methods presented in articles using the same data set conditions.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

[1] Shortliffe, E. H. (1990). Medical Informatics: Computer Applications in Medicine. Addison-Wesley.

[2] Botstein, D., & Risch, N. (2003). Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nature Genetics, 33 Suppl(S3), 228–237. https://doi.org/10.1038/ng1090

[3] Huang, Y., Mccullagh, P., Black, N., & Harper, R. (2005). Feature selection and classification model construction on type 2 diabetic patient's data, Adv. Adv. Data Min, 153–162. https://doi.org/10.1007/978-3-540-30185-1_17

[4] Tama, B. A., Rodiyatul, & Hermansyah, H. (2011). An early detection method of type-2 diabetes mellitus in public hospital. TELKOMNIKA (Telecommunication Computing Electronics and Control), 9(2), 287. https://doi.org/10.12928/telkomnika.v9i2.699

[5] Wolpert, D. H. (1996). The lack of A Priori distinctions between learning algorithms. Neural Computation, 8(7), 1341–1390. https://doi.org/10.1162/neco.1996.8.7.1341

[6] Fang, H., Rizzo, M. L., Wang, H., Espy, K. A., & Wang, Z. (2010). A new nonlinear classifier with a penalized signed fuzzy measure using effective genetic algorithm. Pattern recognition, 43(4), 1393-1401. https://doi.org/10.1016/j.patcog.2009.10.006

[7] Khan, N., Gaurav, D., & Kandl, T. (2013). Performance evaluation of Levenberg-Marquardt technique in error reduction for diabetes condition classification. Procedia Computer Science, 18, 2629–2637. https://doi.org/10.1016/j.procs.2013.05.455

[8] Nai-arun, N., & Moungmai, R. (2015). Comparison of classifiers for the risk of diabetes prediction. Procedia Computer Science, 69, 132–142. https://doi.org/10.1016/j.procs.2015.10.014

[9] Perveen, S., Shahbaz, M., Guergachi, A., & Keshavjee, K. (2016). Performance analysis of data mining classification techniques to predict diabetes. Procedia Computer Science, 82, 115–121. https://doi.org/10.1016/j.procs.2016.04.016

[10] Kandhasamy, J. P., & Balamurali, S. (2015). Performance analysis of classifier models to predict diabetes mellitus. Procedia Computer Science, 47, 45–51. https://doi.org/10.1016/j.procs.2015.03.182

[11] Çalişir, D., & Doğantekin, E. (2011). An automatic diabetes diagnosis system based on LDA-Wavelet Support Vector Machine Classifier. Expert Systems with Applications, 38(7), 8311–8315. https://doi.org/10.1016/j.eswa.2011.01.017

[12] Ganji, M. F., & Abadeh, M. S. (2011). A fuzzy classification system based on Ant Colony Optimization for diabetes disease diagnosis. Expert Systems with Applications, 38(12), 14650–14659. https://doi.org/10.1016/j.eswa.2011.05.018

[13] Upadhyaya, S., Farahmand, K., & 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(15), 5830–5838. https://doi.org/10.1016/j.eswa.2013.05.012

[14] Zhu, J., Xie, Q., & Zheng, K. (2015). An improved early detection method of type-2 diabetes mellitus using multiple classifier system. Information Sciences, 292, 1–14. https://doi.org/10.1016/j.ins.2014.08.056

[15] Beloufa, F., & Chikh, M. A. (2013). Design of fuzzy classifier for diabetes disease using Modified Artificial Bee Colony algorithm. Computer Methods and Programs in Biomedicine, 112(1), 92–103. https://doi.org/10.1016/j.cmpb.2013.07.009

[16] Sharma, M. (2016). Diabetes Prediction by using Bacterial Foraging Optimization Algorithm and Artificial Neural Network. International Journal of Computer Science and Information Technology & Security (IJCSITS), 6.

[17] Swapna, Kp, S., & Vinayakumar. (2018). Automated detection of diabetes using CNN and CNN-LSTM network and heart rate signals. Procedia Computer Science, 132, 1253–1262. https://doi.org/10.1016/j.procs.2018.05.041

[18] Ramezani, R., Maadi, M., & Khatami, S. M. (2018). A novel hybrid intelligent system with missing value imputation for diabetes diagnosis. Alexandria Engineering Journal, 57(3), 1883–1891. https://doi.org/10.1016/j.aej.2017.03.043

[19] Gujral, S. (2017). Early Diabetes Detection using Machine Learning: A Review. IJIRST-International Journal for Innovative Research in Science &Technology, 3.

[20] Fatima, M., & Pasha, M. (2017). Survey of machine learning algorithms for disease diagnostic. Journal of Intelligent Learning Systems and Applications, 09(01), 1–16. https://doi.org/10.4236/jilsa.2017.91001

[21] Somasundaram, N., & Ayyasamy, B. (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(103748), 103748. https://doi.org/10.1016/j.bspc.2022.103748

[22] Suyanto, S., Meliana, S., Wahyuningrum, T., & Khomsah, S. (2022). A new nearest neighbor-based framework for diabetes detection. Expert Systems with Applications, 199(116857), 116857. https://doi.org/10.1016/j.eswa.2022.116857

[23] Prabha, A., Yadav, J., Rani, A., & Singh, V. (2021). Design of intelligent diabetes mellitus detection system using hybrid feature selection based XGBoost classifier. Computers in Biology and Medicine, 136(104664), 104664. https://doi.org/10.1016/j.compbiomed.2021.104664

[24] Mehedi Hassan, M., Mollick, S., & Yasmin, F. (2022). An unsupervised cluster-based feature grouping model for early diabetes detection. Healthcare Analytics, 2(100112), 100112. https://doi.org/10.1016/j.health.2022.100112

[25] Fang, J., Xie, Z., Cheng, H., Fan, B., Xu, H., & Li, P. (2022). Anomaly detection of diabetes data based on hierarchical clustering and CNN. Procedia Computer Science, 199, 71–78. https://doi.org/10.1016/j.procs.2022.01.010

[26] Ordas, T. G., Nbenavides, M., Benitez-Andrades, A., & Alaiz-Moreton, J. (2021). Diabetes detection using deep learning techniques with oversampling and feature augmentation. Computer Methods and Programs in Biomedicine.