Dandao Xuebao/Journal of Ballistics
ISSN: 1004-499X
Vol. 36 No. 1(2024)

Enhancing Diabetes Diagnosis: A Comparative Analysis of Machine Learning Algorithms and Evalution Metrics

1- Mohammad Ordouei*, 2- Bahareh Jalali, 3- Azar Tolouee

1-Department of Computer,South Tehran Branch,Islamic Azad University,Tehran Iran,
Dr.ordouei@gmail.com, ORCID ID: 0000-0002-0391-0638 (Corresponding author)
2-Department of Computer,South Tehran Branch,Islamic Azad University,Tehran Iran,
j.baharehjalali@gmail.com
3-Department of Electrical and Computer Engineering, Ryerson University, Canada,
a.tolouee@gmail.com, ORCID ID: 0009-0009-4172-1617

Abstract: In recent years, artificial intelligence methods have been widely employed in various fields, including medical diagnosis applications. The main objective of this research is to implement a medical decision support system based on fuzzy decision trees and the PSO algorithm for diabetes diagnosis. To achieve this goal, a group-wise classification method is initially used to determine regression coefficients, which are then optimized using the particle swarm optimization (PSO) method. The PIMA dataset available on the UCI website is used to obtain the dataset. Additionally, MATLAB software is utilized for result simulation, and the results of the proposed method are compared with group-wise classification of numerical data, PSO optimization algorithm, and some algorithms used in the research background. The comparison metric is accuracy. The result obtained from applying the proposed method on the dataset used in this study yields an accuracy of 0.9033, indicating the superiority of the proposed method over GMDH and PSO algorithms, as well as other compared methods.

Keywords: Diabetes Disease - Diabetes Diagnosis - GMDH Algorithm - PSO Algorithm.

Introduction:

Disease is an integral part of human life; nowadays, many people become ill, and many are under medical care. Due to the increasing number of patients, the discovery of new diseases, and other factors such as human error, the use of intelligent and automated applications has emerged as a significant option. Medical diagnosis is an area where computer and computational intelligence applications have received much attention. Depending on the patient's condition, the type of disease, as well as the economic and social status of the patient, various methods can be considered for treatment. Consequently, the selection of treatment type and method involves numerous variables, requiring considerable time and high accuracy for decision-making.

Most medical and treatment centers collect various data regarding different diseases and factors contributing to the onset of such diseases. Exploring these data and extracting useful patterns from them is one of the primary objectives of collecting such data.

In recent years, artificial intelligence methods have been widely used in various fields, including medical diagnosis applications. The main objective of this research focuses on issues related to therapeutic care in the field of healthcare. Therapeutic care includes all activities leading to disease diagnosis, treatment, and returning the patient to their pre-disease condition. Clearly, the most important part of treating any disease is the diagnosis stage, which patients undergo, and most efforts made so far using intelligent techniques and methods focus on this aspect.

Diabetes is a common disease that develops at different ages. A diabetic person either lacks insulin production or is resistant to insulin. A diabetic individual is prone to diseases such as heart disease and kidney failure. Due to unhealthy lifestyles, work culture, and lack of physical activity, diabetes has increased at a rapid rate. Although diabetes is a chronic disease, early detection and monitoring prevent its harmful effects. The global prevalence of diabetes and its associated complications is rapidly increasing. Therefore, proper diagnosis and classification of

diabetes are essential to reduce the risk or progression of long-term complications. Nowadays, modern data mining-based methods are used in medical sciences for prediction, early diagnosis, increased accuracy, and obtaining useful results for diabetes prediction. In this research, we intend to focus on diabetes diagnosis based on the GMDH algorithm.

Problem Statement:

When the body cannot produce enough insulin and the glucose level in the blood increases, diabetes develops. Generally, the global impact of diabetes has increased in recent years, and it is expected that this trend will continue soon. Diabetes affects about 463 million people worldwide. By 2045, this figure is expected to increase to 700 million. In many countries, the number of people with type 2 diabetes is increasing. Diabetes has become one of the most common causes of death, leading to major health problems such as blindness, kidney disease, stroke, heart disease, etc.

Diabetes is one of the most common metabolic diseases in Iran and the fifth leading cause of death worldwide. The prevalence of diabetes worldwide has led to the development of new methods in biomedical research, including artificial intelligence. If initial prediction of diabetes is accurate, the severity of diabetes and the risk factor can be significantly reduced. Machine learning, as a result, has gained more popularity in the medical community, especially in the field of diseases.

Importance and Necessity of Research:

Today, in the modern industrial world, the risk of chronic diseases has increased significantly. Among them, diabetes, as the fourth leading cause of death in most developed countries, has received attention and has become one of the most important concerns of the people and officials. Unfortunately, diabetes patients are at risk of developing serious diseases such as cardiovascular, ocular, renal, limb amputation, etc. Therefore, implementing a method that can assist physicians in accurately diagnosing the presence or absence of this disease can be a significant step in preventing and controlling this disease, especially in its early stages. In this regard, various research has been presented, one of which is the use of artificial intelligence algorithms and classification methods. Despite the many research conducted in the field of diagnosing this disease, a definite method has not yet been introduced, and given the increasing development of data mining-based techniques, continuing research in this area and presenting an intelligent method for early and accurate diagnosis seems important and necessary. Based on this, in this research, diabetes diagnosis using the GMDH algorithm and the PSO algorithm has been addressed.

Structure of the Research:

Diabetes, especially in recent times, has become a global health issue as it belongs to a group of metabolic diseases characterized by high blood sugar levels for extended periods, resulting from long-term deficiencies in insulin secretion or function, or both. The metabolic system of the human body and most animals is designed to maintain blood sugar levels using insulin released into the bloodstream by the pancreas. However, in some cases, the pancreas lacks the capacity to produce insulin in sufficient quantities to assist in glucose metabolism in the blood. In most cases, insulin is produced in amounts less than required or insulin has deficiencies in its function, ultimately leading to type 2 diabetes, the most common type of diabetes. This condition affects many organs of the body, including the heart, kidneys, eyes, and nerves, but is not limited to them. While some individuals with type 2 diabetes can manage their blood glucose levels through exercise and healthy eating, others may need medication or insulin to help manage diabetes. Early diagnosis and subsequent appropriate treatment significantly reduce the risk of potential complications [1,2]. The most common symptoms of diabetes include frequent urination, excessive fatigue, feeling hungry even while eating, blurred vision, slow healing of bruises or cuts, weight loss (in type 1), pain, numbness, or tingling in the hands and feet (type 2).

Advancements in the analysis of healthcare can assist both physicians and patients. Healthcare analysis can help identify and diagnose diseases early. Therefore, they can also be used to improve the quality of healthcare and patient outcomes. Machine learning models can be used to find patterns in data and make predictions based on these patterns. They are used in healthcare programs for diagnosis, prediction, and treatment of diseases. With the

development of new algorithms and other technological innovations, these models have become more effective in providing patient care than ever before [3,4].

Literature review

Today, most medical centers collect various data on various diseases and factors contributing to the onset of such diseases. Exploring these data and extracting useful patterns from them is one of the main objectives of collecting such data. Diabetes is one of the diseases that is important and significant due to its hidden symptoms and heavy costs for individuals and society. Diabetes is recognized as one of the deadliest diseases worldwide. It is a metabolic disorder characterized by high levels of glucose in the blood for a prolonged period in the body, as it cannot be properly utilized. Severe complications associated with diabetes include diabetic ketoacidosis, non-ketotic hyperosmolar coma, cardiovascular disease, stroke, chronic kidney failure, retinopathy, and foot ulcers. There is a significant increase in the number of diabetes patients worldwide, making it a major global health problem. Early diagnosis of diabetes is beneficial for useful treatment and reduces the likelihood of severe associated complications.

In this section, some literature on disease diagnosis using machine learning algorithms has been reviewed. In [5], the Random Forest algorithm was used to diagnose diabetes. To assess the performance of the proposed algorithm for diagnosing diabetes in the mentioned study, a dataset containing 768 samples (patients) with 8 features was used. According to the results of this study, the accuracy of diabetes diagnosis was 99.86%. In [9], an expert neural network system based on the Improved Whale Optimization Algorithm was used for diabetes diagnosis. The PIMA database was used for training and testing the proposed neural network. The results of applying the proposed Improved Whale Optimization Algorithm-based neural network system showed that it could diagnose diabetes with an accuracy of approximately 85%. In [6,7], a method using a multi-layer artificial neural network and the Bee Algorithm for diabetes diagnosis has been proposed. In [8], Multi-Layer Perceptron (MLP) neural network, LVQ neural network, Support Vector Machine (SVM), and K-means clustering method were used for diabetes diagnosis, and the mean square error was calculated. The accuracy of each learning algorithm was 94%, 92%, 96%, and 93%, respectively. The results of the mentioned study indicate that the SVM method performed better than other methods in diagnosing diabetes.

Gestational diabetes is associated with multiple short-term and long-term complications in both the mother and the child. Identifying its risk factors can help in timely diagnosis and prevention of its related complications. In [10], the design and comparison of predictive models for gestational diabetes using artificial intelligence algorithms such as Decision Trees and Artificial Neural Networks were discussed. The research community included 1270 pregnant women covered by healthcare centers in Ahvaz city, among whom 816 were healthy and 454 were diagnosed with gestational diabetes. Sensitivity, specificity, accuracy, and precision were calculated to evaluate the performance of the models. Finally, the AdaBoost classification algorithm was used to enhance the proposed model. After principal component analysis, nine variables were selected for initial modeling. The artificial neural network model achieved an area under the ROC curve and sensitivity of 82.3% and 85.1%, respectively, while the decision tree model obtained an area under the ROC curve and sensitivity of 82.6% and 84%, respectively. After removing variables with less weight and enhancing the proposed model, the area under the ROC curve and sensitivity increased. In [11,12], the roles of various data mining methods such as Support Vector Machines, Decision Trees, etc., in predicting this disease were discussed. Furthermore, for diabetes prediction, an investigation and analysis of important diabetes features such as age and body mass index, the data used, and useful tools were discussed. In [13], the possibility of predicting diabetes using data mining techniques was discussed. The research was descriptive-analytical and cross-sectional, conducted on individuals visiting healthcare centers in Mohammadieh County, Qazvin Province, for diabetes screening. The data were analyzed and compared using the k-Nearest Neighbors Algorithm (k-NN), Decision Trees (DT), and Support Vector Machines (SVM). MATLAB® software, version 8.2, was used for data analysis. Study [14,21] aimed to investigate the types of studies conducted in the field of artificial intelligence and diabetes. The mentioned study was conducted using a systematic review method, and reputable internal databases including Irandoc, Magiran, SID, and the Google Scholar search engine were searched using Persian, separately and in combination, with the keywords of artificial intelligence and diabetes without time limitation until June 20, 2021. A total of 7495 retrieved articles were screened in various stages (removal of duplicate articles (1824), title and abstract screening (5884), and full-text screening (30)), and finally, 20 articles meeting the researchers' criteria were thoroughly examined. Among the retrieved articles, 20 articles met the inclusion criteria, with 16 articles focusing on artificial

intelligence-based methods and 4 articles on the design of new artificial intelligence-based systems. Ten articles focused on the role of artificial intelligence in prediction, 8 articles on diagnosis, and 2 articles on the control and management of diabetes. The majority of articles utilized data mining methods such as artificial neural networks, decision trees, etc. (16 articles), and some studies evaluated and compared artificial intelligence methods' application, accuracy, and sensitivity in diabetes diagnosis and prediction. A systematic review of the articles indicated that the use of data mining methods for diabetes management in Iran has been associated with good progress, but further actions are needed in designing artificial intelligence systems and algorithms and in the field of diabetes control and management.

In [15,16], a system for diabetes diagnosis was presented using data mining techniques and employing a combination of artificial neural networks and the particle swarm optimization algorithm. One of the important features of the proposed method is the use of the standard Pima dataset. In this method, along with training the neural network, the particle swarm optimization algorithm was used to determine the optimal weights of the neural network to construct an accurate diabetes prediction model. The proposed method was evaluated with three reliable techniques for diabetes diagnosis, including regression, artificial neural network, and decision tree, based on accuracy, specificity, and sensitivity criteria. Simulation results showed that the proposed method outperformed in all three performance criteria, being highly consistent with the real model. The highest values of accuracy, specificity, and sensitivity in the proposed method with 50 different experiments were 94.1%, 92.88%, and 92.12%, respectively.

In [17,20,23], diabetes diagnosis was addressed. This study involved 8 diabetic patients and 64 healthy individuals. Electrocardiography was performed for all subjects, and the required information from ECG images, including patient name, age, HR, p, t, RR, PP, P, PR, qt, and qtcb, was extracted and collected in the database. Standard data mining algorithms and neural networks were used for patient classification. Data were analyzed and evaluated using various data mining algorithms and classification techniques, and the results of each were compared based on the correct rate. Weka software was used for classifications. In this study, the accuracy of identification using rule-based algorithms and neural networks showed better results in diabetes diagnosis compared to decision tree algorithms and distance-based algorithms.

Thus far, artificial intelligence technologies have been widely used in medicine and healthcare. However, new artificial intelligence technologies are becoming increasingly complex, making understanding them difficult for users. Problems in understanding artificial intelligence technologies can lead to users' reluctance to trust, use, or recommend related applications. To overcome this problem, the concept of eXplainable Artificial Intelligence (XAI) has been proposed. XAI aims to increase the usability of existing artificial intelligence technology by explaining the execution process and results. Explainable Artificial Intelligence (XAI) tools are used to increase the usability of artificial intelligence technologies by explaining their processes and results. In most past research, XAI tools and techniques were usually only applied to the inference part of AI programs. In [18,19,22], a systematic approach to increase the explainability of AI applications in healthcare was proposed. To demonstrate the usefulness of the proposed method, several AI applications for type 2 diabetes diagnosis were considered as examples. According to experimental results, XAI tools and technologies in the proposed method were more diverse than past research. Additionally, in the mentioned study, an artificial neural network was approximated to simpler and more intuitive classification, and the Classification and Regression Tree (CART) was used with the explanation of the interpretable local model (LIME), and the extracted rules were used to recommend users to restore their health.

In [13], the definition and classification of diabetes were discussed, and various diagnostic criteria were explained.

Methodology

Information Sources and Data Collection Methods

In this research, library research method is used to develop foundations, definitions, theoretical concepts, literature, and research background, with the most important sources being the internet and the libraries of national universities. The UCI website was used to obtain the required dataset.

Research Method

With the rapid growth of artificial intelligence technology and its integration into medicine, detailed datasets for diagnosis are very useful, especially as they ensure that trends in the dataset are decodable using data mining techniques. Such information is beneficial for healthcare providers for more accurate diagnosis and increased awareness of health conditions where detailed data is available. In the method presented in this research, the GMDH algorithm is used for diabetes diagnosis.

In this research, we aim to prepare and preprocess the data for modeling after understanding the data. In the initial modeling step, an appropriate method must be selected. In this study, we have used a method based on the GMDH algorithm for modeling. Patient features are entered into Excel, preprocessed using preprocessing methods, and prepared for entry into the desired algorithm. Then, using the implementation of combined algorithms in MATLAB software, patient feature patterns are extracted, evaluated, analyzed, and the results are scrutinized.

Dataset

For the present research, the PIMA dataset available on the UCI website was used.

The PIMA diabetes dataset consists of medical examination data related to the PIMA community women in Phoenix, Arizona, USA. Given the highest prevalence of type 2 diabetes, they have been the subject of research studies. This dataset contains medical examination data of 768 individuals, including 500 samples of non-diabetic individuals and 268 samples of diabetic individuals, with features related to diabetic patients. The PIMA dataset is a binary class label dataset, with "1" indicating a positive result and "0" indicating a negative result.

1. Variables in this dataset include: Number of times pregnant
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/(height in m)^2)
7. Diabetes pedigree function
8. Age (years)
9. Class variable (0 or 1)

The desired method for data analysis in this research is the combined algorithm of GMDH and PSO, with MATLAB software being used. The proposed method in this study is as follows:

1. Data Preparation: Inputs are represented as $x2,...,xm$ and each sample includes features represented as $x12,...,x1n$ where $x11=x1$ is visible.
2. Data Preprocessing: Data must be standardized in all aspects for effective use and better results. In this section, data are thoroughly examined from multiple aspects, with full attention given to them. Before connecting the dataset to the proposed model, preprocessing is performed, including normalization, handling unknown data, feature selection, etc., to ensure the dataset's suitability.
3. Handling Unknown Data: Lost data are recognized as an important issue to be carefully examined in the preprocessing stage before applying machine learning algorithms to achieve effective results. In practice, a dataset may contain unknown data due to various reasons such as human errors, equipment malfunctions, data unavailability, or data inconsistency. Different approaches exist to solve this problem, including removing

samples with unknown data, replacing unknown data with averages, medians, global constants, and other methods that are used to replace unknown data.

4. Normalization: In the preprocessing stage, to obtain better results, the values of each feature are normalized between 0 and 1. Normalization is essential for better results. The normalization of each dataset is done using the equation (1), where *Xmax* and *Xmin* represent the maximum and minimum values in the domain of attribute *X*. After normalization, the values of all attributes are placed in the range of [0, 1].

$$\text{Normalize}(x) = \frac{(\text{x} - \text{X}_{\min})}{(\text{X}_{\max} - \text{X}_{\min})} \quad (1)$$

Data Division:

To apply machine learning algorithms effectively, it is necessary to use data that were not involved in the training process and are considered new data for the machine; otherwise, the evaluation would be meaningless. For this purpose, data are divided so that a portion of the data is set aside as test data and not used in the training process.

Classification Using the Combined Method of GMDH and PSO:

Initially, numerical data are categorized by a group classification model into a nonlinear regression relationship that represents the output-input relationships. Then, using the particle swarm optimization algorithm, the coefficients are improved. In this way, classification with higher accuracy is achieved. The group classification of numerical data is a type of neural network model that derives the relationship between inputs and outputs in the form of polynomials. The relationship between output parameters and inputs using the complex discrete Volterra series form (2) is formulated below:

$$\text{y}=a_0 + \sum_{i=1}^{n} a_i x_i + \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} x_i x_{j+} \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} a_{ijk} x_i x_j x_k + \cdots \quad (2)$$

In many practical applications, the second-degree and two-variable forms of these polynomials are used in the form of equation (3):

$$\hat{y} = G(x_i, x_j) = a_0 + a_1 x_i + a_2 x_j + a_3 x_i^2 + a_4 x_j^2 + a_5 x_i x_j \quad (3)$$

The unknown coefficients a_i in equation (3) are determined such that the difference between the actual output and the observed output is minimized. In this study, we aim to first obtain the regression coefficients of the above equation using the group classification method for numerical data and then consider these coefficients as initial values in the particle swarm optimization algorithm to improve them.

Results Evaluation:

For a disease prediction model, the performance of a classifier plays a crucial role. The proposed algorithm of this study is compared with previous research. The criteria considered for assessing the performance of the proposed algorithm in this study include accuracy, sensitivity, and specificity. Accuracy represents the ratio of correctly labeled positive and negative cases. Sensitivity indicates the ratio of positive cases correctly labeled as positive. Specificity represents the ratio of negative cases correctly labeled as negative. In this study, positive cases refer to individuals without diabetes, and negative cases refer to individuals with diabetes.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (4)$$

TP is the number of samples belonging to the "individuals with diabetes" class and predicted as "individuals with diabetes" by the algorithm.

FP is the number of samples belonging to the "individuals without diabetes" class and predicted as "individuals with diabetes" by the algorithm.

Execution of the Proposed Method

Preprocessing:

The PIMA dataset comprises 768 data points with 9 attributes. This dataset includes medical data from 768 patients with 8 different test features. The ninth attribute is the class attribute indicating the diagnosis outcome. Six features, including Pregnancy, Glucose, Blood Pressure, Skin Thickness, Insulin, and BMI, have zero values for some samples in the dataset, with frequencies of 111, 5, 35, 227, 374, and 11, respectively. Most of these zero values might result from data collection errors, typographical errors, or missing data during the data collection process. These zero values might negatively affect the classifier's performance. In the present study, all rows containing unknown values were removed. The remaining records amount to 393, consisting of 130 positive samples and 263 negative samples.

Normalization

Upon inspecting the data, it was found that there are no duplicate values in the dataset. The min-max normalization method was used for normalization.

Data Division

In many predictive modeling tasks, 80% of the dataset is commonly used for training purposes. Similarly, in this study, 80% of the total dataset was used as the training set, and the rest was used as the test set.

Classification:

Application of the GMDH Algorithm

After preprocessing the data, the group classification algorithm for numerical data was applied to the resulting dataset. In this algorithm, the maximum neuron value in each layer and the maximum number of layers were determined using a trial-and-error approach. Various combinations of values were tested, and the combination of 20 and 5 yielded the best results for the maximum neuron value per layer and the number of layers, respectively. After applying these values, 20 neurons were obtained in the first to fourth layers, and 1 neuron was obtained in the fifth layer. The relationship between input and output variables was specified for each neuron, and finally, they were combined. For example, in the first layer, the first neuron considered variables 9 and 18, and the coefficients were determined according to equation (4). In the last layer, i.e., the fifth layer, variables 3 and 6 were considered, and the coefficients were determined according to equation (4-2). All these parameters were considered in the particle swarm optimization (PSO) algorithm. In all stages of this study, 80% of the dataset was used as the training set, and the remainder was used as the test set.

$$\hat{y} = G(x_i, x_j) = -0.0683 + 0.0105x_9 + 0.5226x_{18} - 0.5244x_9^2 - 0.5226x_{18}^2 - 0.0098x_9x_{18} \quad (4)$$

$\hat{y} = G(x_i, x_j) = 0.0132 + 0.2182x_3 + 0.3747x_6 - 0.5868x_3^2 - 0.3404x_6^2 + 1.3231x_3x_6$	(5)

Application of the Particle Swarm Optimization (PSO) Algorithm

In the previous stage, i.e., applying the GMDH algorithm, regression coefficients were obtained. These coefficients were used as the initial values for the PSO algorithm. A combination of different equations obtained from the GMDH was used as the initial position for the PSO algorithm. Each coefficient obtained from the neurons in GMDH was considered as a particle in PSO. For instance, the coefficients obtained from equations (4) and (5) and similar equations in other layers and neurons were considered. In the PSO algorithm, the number of iterations was set to 100 using a trial-and-error approach, and the number of particles was set to 50. The particles were defined as a function. For example, for the last layer, whose formula is given in equation (5):

```
model.fhat=@(x1,x2,x3,x4,x5,x6,x7,x8,x9,x10,x11,x12,x13,x14,x15,x16,x17,x18,a)
 a(1)+(a(2)*(x3))+(a(3)*(x6))+(a(4)*(x3^2))+(a(5)*(x6^2))+(a(6)*(x3*x6));
```

The initial position was set as follows:

particle(1).Position=[0.0132, 0.2182, 0.3747, 0.5868, 0.3404, 1.3231];

The above values were obtained from GMDH. After applying PSO, the coefficients were improved. For example, for the final coefficients obtained as mentioned above:

2.91155966385184- 1.43446397563053 3.91136841521325- 4.90758917668994 0.0167872701066379 0.569112107528532-

The results of the proposed method were compared with those of both the group classification of numerical data and the PSO optimization algorithm. The particles were compared individually.

The confusion matrix of the training and test datasets was obtained according to tables (1) and (2).

Table 1: Confusion Matrix resulting from the proposed model in this study for the training dataset.

		Predicted Model	
		Individuals without diabetes	Individuals with diabetes
Actual Values	Individuals without diabetes	244	19
	Individuals with diabetes	16	114

Table 2: Confusion Matrix resulting from the proposed model in this study for the test dataset

		Predicted Model	
		Individuals without diabetes	Individuals with diabetes
Actual Values	Individuals without diabetes	243	20
	Individuals with diabetes	18	112

The remaining records amount to 393, consisting of 130 positive samples and 263 negative samples.

In tables (3) and (4), the values of accuracy, sensitivity, and specificity obtained from the proposed method for the training and test datasets are provided.

Table 3: Accuracy values of the proposed method for the training dataset (in percentage)

Accuracy of proposed method
0.9109

Table 4: Accuracy, Feature, and Sensitivity values of the proposed method for the test dataset (in percentage)

Accuracy of proposed method
0.9033

Tables 5 and 6 provide a comparison of the accuracy, sensitivity, and transparency obtained from the proposed method and the compared methods for the training and test dataset.

Table 5: Comparison of the proposed algorithm in this study and PSO and GMDH for the training dataset:

	Accuracy
GMDH	**0.86**
PSO	**0.8**
Proposed Method	**0.9109**

Table 6: Comparison of the proposed algorithm in this study and PSO and GMDH for the test dataset:

	Accuracy
GMDH	**0.70**
PSO	**0.58**
RF	**0.8861**
Proposed Method	**0.9033**

Conclusion

One of the most striking phenomena over the past hundred years has been the replacement and prevalence of chronic diseases such as cardiovascular diseases, hypertension, diabetes, asthma, and allergies instead of acute and infectious diseases. Among them, diabetes is one of the most common diseases in human societies and is considered a major health problem worldwide, posing a significant challenge to community health. Due to its early and late complications, this disease imposes many problems on both the patient and society. Late complications such as eye, kidney, heart, vascular, and neurological complications lead to various disabilities. Diabetes is the most common cause of organ failure, including blindness, kidney failure, and non-traumatic amputation. Moreover, diabetes is a major risk factor for atherosclerosis, which itself is one of the most common causes of death. This disease is on the rise due to industrialization and urbanization. Sedentary lifestyle, dietary patterns, family history, stress, and some environmental and genetic factors are involved in the occurrence of this disease.

Timely diagnosis of diseases, including diabetes, is essential for treatment advancement. In this study, the diagnosis of diabetes using GMDH and PSO was addressed, evaluated using MATLAB software, and compared with the method of numerical data classification and particle swarm optimization algorithm, as well as the random forest algorithm used in previous research. The evaluation criterion considered is accuracy. The proposed algorithm has performed better in terms of accuracy compared to other algorithms evaluated.

References:

[1] Ejiyi, C. J., Qin, Z., Amos, J., Ejiyi, M. B., Nnani, A., Ejiyi, T. U., ... & Okpara, C. (2023). A robust predictive diagnosis model for diabetes mellitus using Shapley-incorporated machine learning algorithms. Healthcare Analytics, 3, 100166.

[2] M. Ordouei and T. BaniRostam, Integrating data mining and knowledge management toimprove customer relationship management in banking industry (Case study of Caspian Credit Institution), Int. J. Comput. Sci. 3 (2018), 208–214.

[3] M. Ordouei and T. Banirostam, Diagnosis of liver fibrosis using RBF neural network and artificial bee colony algorithm, Int. J. Adv. Res. Comput. Commun. Engin. 11 (2022), no. 12, 45–50.

[4] M. Ordouei and M. Moeini, Identification of female infertility in people with thalassemia using neural network, Int. J. Mechatron. Electeric. Comput. Technol. 13 (2023), no. 48, 5371–5374.

[5]. Ordouei, I. Namdar." Web Robot Detection Based On Fuzzy System and PSO Algorithm", IJCSN International Journal of Computer Science and Network, Volume 7, Issue 4, August 2018.

[6] M Ordouei, A Broumandnia, T Banirostam, A Gilani, Providing A Novel Distributed Method For Energy Management In Wireless Sensor Networks Based On The Node Importance Criteria, Journal of Namibian Studies:History Politics Culture, 2023.

[7] A Moradi, M Ordouei, SMR Hashemi, Multi-period generation-transmission expansion planning with an allocation of phase shifter transformers, Int. J. Nonlinear Anal. Appl. In Press, (2023) 1–12.

[8] M. Ordouei, A. Broumandnia, T. Banirostam and A. Gilani, Optimization of energy consumption in mart city using reinforcement learning algorithm, Int. J. Nonlinear Anal. Appl. In Press, (2022) 1–15.

[9] M Ordouei, A Shams, M Moeini, ARTIFICIAL INTELLIGENCE ROUTING ALGORITHMS IN INTER-VEHICLE MOBILE NETWORKS, Vol. 10, Issue 08, pp. 8751-8757, August, 2023.

[10] M Ordouei, B Jalali, AS Nourbakhsh, Proposing a new framework for optimizing energy consumption in sensor nodes used in the Internet of Things, Journal of Systems Engineering and Electronic, Vol. 34, Issue 02, pp. , February 2024.

[11] M Ordouei, A Broumandnia, T Banirostam,A Gilani, Efficient energy management in a smart city based on multi-agent systems over the Internet of Things platform, International Journal of Nonlinear Analysis and Applications, pp. 1-8, January, 2023.

[12] Chang, V., Ganatra, M. A., Hall, K., Golightly, L., & Xu, Q. A. (2022). An assessment of machine learning models and algorithms for early prediction and diagnosis of diabetes using health indicators. Healthcare Analytics, 2, 100118.

[13] Moshrefzadeh, S., Ravaii, B., Koozegar, E. A. (2021), Diagnosis of diabetes using random forest algorithm, Iranian Journal of Diabetes and Metabolism, Volume 21, Number 2.

[14] Ahmadiannejad, F., Azadian, H., Taghizadeh, M. (2020), Improved diagnosis of diabetes based on deep learning and enhanced whale algorithm, Fifth International Conference on Innovation and Research in Engineering Sciences.

[15] Abbasi Dezfuli, M., Mohammadi, A., Pourafshar, M., Gorgin, A. (2018), Diagnosis of diabetes using artificial intelligence algorithms, Second National Conference on Computer, Information Technology and Applications of Artificial Intelligence, Ahvaz.

[16] Abedian, I., Ayoubi, A., Ghaffari, H., Zahabi, I. (2019), Diagnosis of diabetes using data mining-based methods based on native data, Journal of Torbat Heydariyeh University of Medical Sciences, Volume 7, Number 1.

[17] Zarei, J., Izadi, M., Azizi, A., Noh Javad, S. (2022), Early prediction of gestational diabetes using decision tree and artificial neural network algorithms. Journal of Endocrinology and Metabolism of Iran. 1401; 24 (1): 1-11

[18] Yaghoubzadeh, R., Kamel, S., Khiraabadi, M. (2017), A review of data mining methods for the diagnosis of diabetes, National Conference on the Application of Modern Technologies in Science and Engineering, Computer

[19] Nazari, M., Zamani Dehkordi, B., Kiyomarsi Dehkordi, F. (2017), Diagnosis of diabetes based on information extracted from ECG signals using artificial neural networks. Journal of Shahrekord University of Medical Sciences. 19 (4): 64-77

[20] Bahador, F., Sabahi, A., Jalali, S., Amiri, F. (2022). Investigating the role of artificial intelligence in diabetes management in Iran: a systematic review, Payavard Salamat Journal, Volume 16, Issue 6.

[21] Keyani Moein, Z. (2018). Diagnosis of diabetes using data mining technique and neural network, Second International Conference on Electrical Engineering, Computer Science and Information Technology, Hamedan.

[22] Wang, Y. C., Chen, T. C. T., & Chiu, M. C. (2023). A systematic approach to enhance the explainability of artificial intelligence in healthcare with application to diagnosis of diabetes. Healthcare Analytics, 3, 100183.

[23] Jagannathan, R., Tamura, K., & Vellanki, P. (2021). Diabetes mellitus: diagnosis and heterogeneity.