Scientia Iranica D (2022) 29(5), 2394–2404

Sharif University of Technology
Scientia Iranica
Transactions D: Computer Science & Engineering and Electrical Engineering
http://scientiairanica.sharif.edu

Evaluation of the impact of environmental conditions on diabetes using ensemble classifier based on genetic algorithm

F. Khademi[a], H. Motameni[a,*], M. Rabbani[b], and E. Akbari[a]

a. *Department of Computer Engineering, Sari Branch, Islamic Azad University, Sari, Iran.*
b. *Department of Applied Mathematics, Sari Branch, Islamic Azad University, Sari, Iran.*

Received 25 June 2021; received in revised form 13 March 2022; accepted 25 April 2022

KEYWORDS
Medical data mining;
Diabetes;
Environmental conditions;
Ensemble classifier;
Genetic algorithm.

Abstract. Medical data mining is considered as a new solution to the analysis of medical data and further expansion of the relative knowledge. Medical data mining has a high potential to discover the hidden patterns. Recently, some studies have been conducted on the relationship between the individuals' environmental surroundings and their exposure to diseases, thus clearly confirming the considerable impact of the quality of environmental indicators such as environmental pollutants on diseases. In this respect, the current study investigated the effect of the environmental conditions on the chance of diabetes based on medical data mining technique. Given that diabetes is a global threat to the human health, an Ensemble Classifier based on Genetic Algorithm (ECGA) method was designed to study the environmental conditions in diabetes. Decision tree, random forest, K-nearest neighbor, and naive were included in the designed ensemble classifier. It was found that ECGA was more accurate than the base classifier algorithms. For study purposes, three datasets were collected from different regions in Iran with different climatic conditions and it was found that environmental conditions had a significant impact on diabetes.

1. Introduction

Today, diseases highly contribute to increasing the mortality rates. In this regard, medical data mining provides solutions to this problem by analyzing a medical dataset. It is also regarded as an efficient analytical method for extracting the required information from large volumes of medical data. Data mining techniques are used to develop predictive models that are able to classify and predict the medical datasets [1]. Diabetes is considered as one of the most important health challenges worldwide. It is associated with sudden abnormal increase in the blood glucose levels which is generally found in two types. While Type 1 diabetes is caused by insufficient production of insulin in the body, Type 2 diabetes is caused by lack of cells in the body's effective response to insulin. This disease causes serious damages to the vital systems of the human body, especially the nervous one. The number of diabetics worldwide is rapidly increasing, and it is estimated that 642 million people will be affected by the disease by 2040 [2]. Figure 1 shows the growth rate of the number of diabetic patients worldwide based

*. *Corresponding author. Fax: +981134445112*
E-mail addresses: f_khademi2010@yahoo.com (F. Khademi); motameni@iausari.ac.ir (H. Motameni); mrabbani@iausari.ac.ir (M. Rabbani); ebrahimakbari30@yahoo.com (E. Akbari)

doi: 10.24200/sci.2022.58499.5758

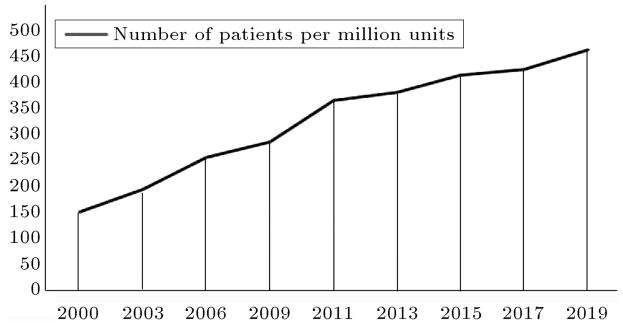

Figure 1. The growth rate of the number of diabetic patients worldwide.

on the reports provided by the International Diabetes Federation [3–10].

Surprisingly, it has been recently found that diabetes and climate change are linked. To be specific, severe climate changes and subsequent rising temperatures may increase the morbidity and mortality rates among patients with diabetes, especially those with cardiovascular complications [11]. The present study aims to evaluate the effects of the environmental conditions on diabetics based on medical data mining techniques. To this end, Ensemble Classifier based on Genetic Algorithm (ECGA) method was designed to investigate the environmental conditions and their effects on diabetes. In addition, decision tree, random forest, K-nearest neighbor, and naive Bayes were used in the designed ensemble classifier of the expert systems. Examination of the environmental conditions based on this method facilitates studying other diseases such as the Coronavirus pandemic. Coronavirus (COVID-19) is the descendent of acute Coronavirus 2 syndrome (SARSCOV2) which turned into an unprecedented global health crisis since 2020. Considering the worldwide epidemic of COVID-19, this method helps examine the environmental impacts on different factors such as the pandemic mortality rates based on real data including physiological characteristics and treatment outcome of the patients with COVID-19 in different parts of the world. However, this issue should be further explored in the future due to the lack of available sufficient data at the moment. This study pursues two main objectives:

1. Designing an the ECGA to increase the accuracy of diabetes diagnosis;
2. Investigating the impact of environmental conditions on diabetes using the designed ECGA in this paper.

The rest of this paper is organized as follows. Section 2 introduces the related classification techniques in two groups of patients with diabetes and potential diabetes. Section 3 describes the methodology of the proposed approach. Section 4 presents the experimental results. Finally, Section 5 concludes the study with some directions for future research.

2. Related work

So far, numerous methods have been proposed in the literature to detect and diagnose diabetes based on the medical data mining techniques. For instance, Huang et al. proposed a classification model to diagnose Type 2 diabetes in patients. To this end, first, they collected the required information about the diabetic patients between 2000 and 2004 from some specific hospitals. Then, they employed the feature selection technique to improve the computational efficiency. Followed by selecting the appropriate features, they utilized three different classifiers (i.e., Nave Bayesian, IB1, and C4.5) to predict the probable conditions of diabetic patients [12]. Kahramanli and Allahverdi, in another study, established a combined system based on artificial and fuzzy neural networks to efficiently diagnose diabetes. The issue of real-time performance was examined to determine the feasibility of the hybrid system [13]. Temurtas et al. conducted a comparative study on the neural network techniques adopted to identify diabetic patients. The structures of both multilayer and possible neural networks were compared to identify diabetic patients [14]. In addition, Patil et al. proposed a hybrid predictive model for Type 2 diabetes. In this prediction model, Cummins clustering algorithm was used to validate the classes assigned to the data while the C4.5 algorithm was applied for classifying the dataset [15]. Moreover, Ganji and Abadeh presented a fuzzy classification system based on ant colony optimization algorithm to diagnose diabetes. This classification system was designed to derive a set of fuzzy rules for diagnosing diabetes [16]. Çalişir and Dogantekin introduced an automatic diabetes diagnosis system based on linear discriminant analysis and Morlet wavelet support vector classifier. In their proposed system, linear separator analysis was used to extract the features and reduce them as well, and Morlet wavelet Support Vector Machine (SVM) was used to classify the dataset [17]. Aslam et al. considered genetic programming in production of new diabetes features. Of note, genetic programming produces new diabetic features by making nonlinear combinations of key features [18]. Seera and Lim proposed a hybrid intelligent system to classify the medical data as diabetic data. The components of this intelligent system were the fuzzy minimum-maximum neural network, regression tree, and random forest [19]. Gorzałczany and Rudziński used a fuzzy classification system based on a multi-objective genetic algorithm for disease diagnosis. In this method, the multi-objective genetic algorithm selects the best rules for making decisions on the diseases [20]. Kavakiotis et al. presented a

systematic review of machine learning and data mining techniques in the field of diabetes. In their efforts, they mainly focused on reviewing the already conducted studies on the diagnosis, prediction, and detection of disease complications [21]. Kumari et al. employed the machine learning algorithms to ensure better accuracy while predicting diabetes. They also used the Pima Indians diabetes dataset in their research. In their proposed method, they used an ensemble of three random forest, naive Bayes, and logical regression algorithms. Compared to the basic ensemble classifier such as AdaBoost, XGBoost, etc., the proposed method enjoyed the benefit of better accuracy [22]. Zheng et al. established a framework based on machine learning for diagnosing diabetes Type 2 through electronic health records. The proposed framework contains the K-nearest neighbor, naive bayes, decision tree, random forest, SVM, and logistic regression [23]. Gupta et al. developed a computational approach to diagnosing and predicting diabetes which was comprised of 15 different machine learning methods such as Perceptron, K-nearest neighbors, SVM, logistic regression, nave bayes, decision tree, random forest, etc. They used two clinical datasets and confirmed the superiority of the logistic regression in the accuracy parameter over other classifications [24]. In another study, Rahman et al. proposed a deep learning method based on Convolutional LSTM (Conv-LSTM) to diagnose diabetes. For this purpose, they utilized the pima Indians diabetes dataset. The obtained results were compared with those from the three methods namely Convolutional Neural Network (CNN), traditional LSTM, and CNN LSTM, indicating that Conv LSTM was superior to other three models in terms of accuracy [25]. Badiuzzaman Pranto et al. used machine learning techniques to predict diabetes. The datasets they used included both Pima Indians diabetes and hospital patient data in Bangladesh. First, they performed initial preprocessing of the dataset and then, applied four categories of decision tree, random forest, K-nearest neighbor, and nave Bayes to the dataset. In both datasets, random forest and nave Bayes outperformed other categories in terms of accuracy [26]. Canadasan Kannadasan et al. proposed a classification model for diabetes Type 2 data based on automated coding of deep neural networks. In this respect, they extracted appropriate features from the dataset through the automated coding and then, used deep neural networks for classification [27]. Patra et al. suggested an improved neighbor classification to predict diabetes. In the classic KNN classifier, Euclidean distance technique was used to calculate the distance between the test and training samples. Of note, in the improved KNN classifier, a new approach to calculating the distance between the test and training samples was proposed based on the standard deviation. The diagnosis accuracy of the proposed approach with regard to the standard dataset of diabetes was 83.20% [28]. Khanam and Foo provided a comprehensive assessment of machine learning algorithms for diagnosing diabetes. In this method, seven machine learning algorithms were evaluated to diagnose diabetes and finally, a neural network model with two hidden layers obtained the highest diagnostic accuracy of 88.60% [29]. Kaul and Kumar provided a comprehensive overview of the studies conducted to develop a model for predicting diabetes. Different machine learning classification algorithms were used to diagnose diabetes. In order to conduct the current study, different machine learning classification algorithms namely the genetic algorithm, decision tree, random forest, logistic regression, SVM, and Naive Bayes were thoroughly investigated. Then, the tests were performed on the Pima Indian Diabetes Dataset (PIDD) found in the UCI machine learning repository. The findings revealed that the best classification performance was exhibited using genetic algorithm [30]. Sangien et al. predicted diabetes using classification algorithms. To this end, they employed three common predictive algorithms namely the SVM, logistic regression, and random forest and diagnosed diabetes using the PIDD. They found that SVM was obtained with the highest accuracy of 80% compared to the other two algorithms [31].

3. Methodology

This study pursued two main objectives: first, designing an ECGA to increase the accuracy of diabetes diagnosis and second, evaluating the environmental impacts on diabetes through an ECGA. This section comprises the following subsections: Data collection, ensemble classifier, and environmental conditions.

3.1. Data collection

Three datasets were collected from health centers in three regions with different climates in Iran to achieve the main objectives listed below:

1. The dataset collected from the first region is represented by DB_1. Based on the DB_1 dataset, $Train_1$ and $Test_1$ datasets are obtained;
2. The dataset collected from the second region is denoted by DB_2. Based on the DB_2 dataset, $Train_2$ and $Test_2$ datasets are obtained;
3. The dataset collected from the third region is represented by DB_3. Based on the DB_2 dataset, $Train_3$ and $Test_3$ datasets are obtained;
4. The integrated form of the three datasets of DB_1, DB_2, and DB_3 is represented by DB. Based on the DB dataset, $Train$ and $Test$ datasets are obtained.

3.2. Ensemble classifier

The designed ensemble classifier consists of expert systems of the decision tree, random forest, K-nearest neighbor, and naive Bayes. In the designed ensemble classifier, ten expert systems were used to diagnose diabetes. In addition, a weight was assigned to each expert system, indicating the extent to which the expert system affects the decision of the ensemble classifier. The more the weight intended for an expert system in the ensemble classifier, the greater the influence of that expert system in the decision-making of the ensemble classifier, and vice versa. The genetic algorithm is used to properly weigh the expert systems in the ensemble classifier. Figure 2 shows the framework of the ensemble classifier designed in this paper.

Table 1 shows the expert systems used in the ECGA. In this group, the test specimens are effective in classifying each expert system. The test samples have two classes, i.e., zero and one, where class zero represents the healthy samples and class one the patient samples. GEC is evaluated upon applying it to the DB dataset.

Chromosome creation

Each chromosome is made up of ten genes, each representing the weight assigned to an expert system in the ensemble classifier. The weight considered for each expert system in the ensemble classifier has a numerical value of 0–1. Figure 3 shows an example of a chromosome.

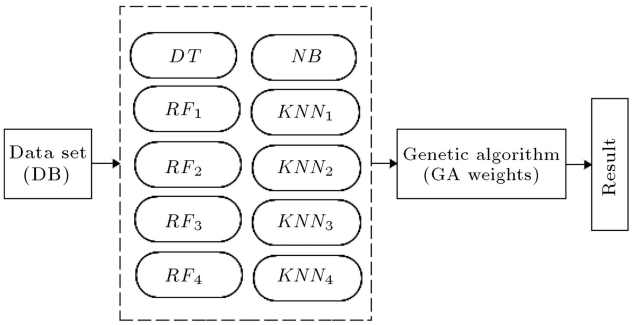

Figure 2. The designed ensemble classifier framework.

Table 1. The expert systems used in the genetic-based ensemble classifier.

No.	Expert system	Abbreviation	Exclusivity
1	Decision Tree	DT	—
2	Random Forest	RF_1	100 Tree
3	Random Forest	RF_2	150 Tree
4	Random Forest	RF_3	200 Tree
5	Random Forest	RF_4	250 Tree
6	Nave Bayesian	NB	—
7	K-Nearest Neighbors	KNN_1	$K = 1$
8	K-Nearest Neighbors	KNN_2	$K = 3$
9	K-Nearest Neighbors	KNN_3	$K = 5$
10	K-Nearest Neighbors	KNN_4	$K = 7$

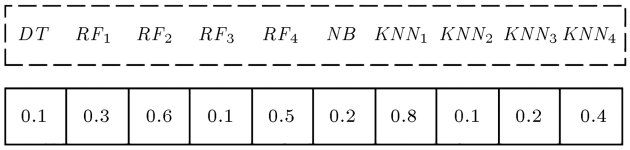

Figure 3. An instance of a chromosome.

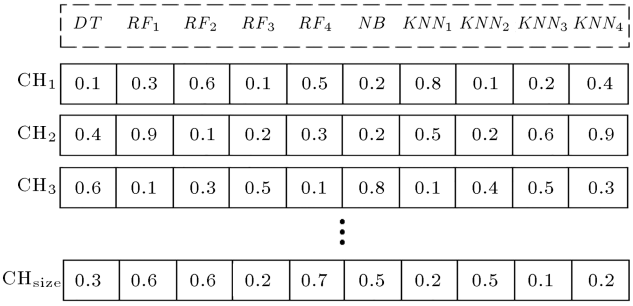

Figure 4. The initial population of the genetic algorithm.

Initial population

The initial population of the genetic algorithm is generated randomly. The genes on each chromosome in the initial population randomly receive a value between 0 and 1. The population size is displayed in the genetic algorithm with the variable size. Figure 4 shows the initial population of the genetic algorithm.

Fitness function

The fitness function is used to evaluate the quality of each chromosome. As mentioned earlier, each chromosome comprises ten genes, each of which represents the weight intended for the expert systems. A test sample is $+1$ or -1 for each expert system. If the test sample class is detected as healthy for an expert system, the number $+1$ will be considered for it. On the contrary, if the test sample class is detected as sick for an expert system, the number -1 will be allocated to it. Finally, ten expert systems were calculated through Eq. (1), as shown below:

$$Final\_Decision = sgn\left(\sum_{i=1}^{10}(c_i * y_i)\right), \quad (1)$$

where y_i represents the ith expert system and c_i the weight of the ith expert system derived from the chromosome data. In addition, the sgn function, called the sign function, is the one which generates three numeric values of $+1$, -1, and 0 to the output. In case the sgn function returns zero as the output, the decision will fail. According to Eq. (1), the final decision of the ensemble classifier is calculated based on the information obtained from the chromosome on

all test samples. Measurement of diagnostic accuracy is considered as a fitness function. In Eq. (2), the diagnostic accuracy criteria are shown as:

$$Fitness = \frac{TP + TN}{TP + TN + FP + FN}, \tag{2}$$

where TP and TN variables stand for the numbers of healthy and patient specimens, respectively, which were diagnosed correctly. In addition, the FP and FN variables indicate the number of healthy and sick specimen, respectively, which were misdiagnosed.

Selection operator

The selection operator is responsible for generating the next generation in the genetic algorithm. The genetic algorithm uses the roulette wheel selection operator. It operates based on the probability of selecting chromosomes. The probability of selecting each chromosome is calculated through Eq. (3). Then, Figure 5 shows the roulette wheel selection operator on four chromosomes.

$$SP(CH_i) = \frac{Fitness(CH_i)}{\sum_{j=1}^{size}(Fitness(CH_i))}. \tag{3}$$

Crossover operator

To apply the crossover operator, a random number is generated in the range of 0–1. In case this number is less than the value specified for P_c, the crossover operation is made between the two parent chromosomes; otherwise, the two-parent chromosomes will be passed to the next step unchanged. The two-point crossover technique is used to execute the crossover operator. In such an operator, two random numbers in the range of "one to ten" are first selected. The numbers selected in this range are the same as the crossover points.

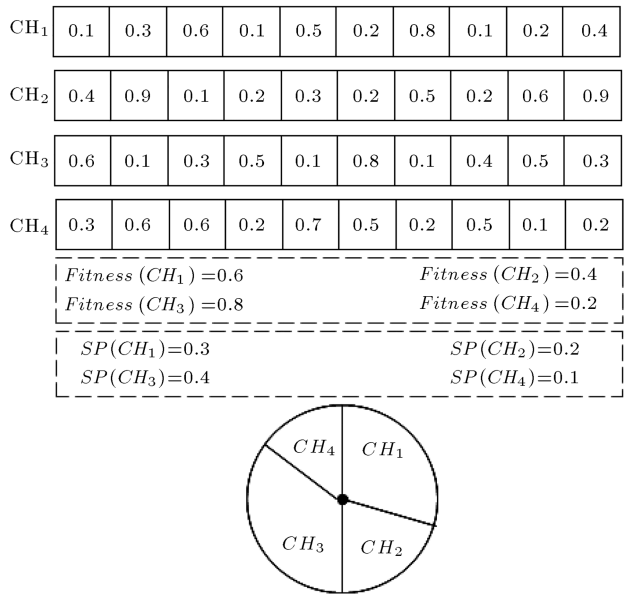

Figure 5. The roulette wheel selection operator.

After selecting the crossover points on the two-parent chromosomes, the genes between these points on the chromosomes move together to produce two pieces of new offspring.

Mutation operator

A random number is generated in the range of 0–1 to apply the mutation operator. If this number is less than the value specified for P_m, the mutation occurs on the parent chromosome; otherwise, the parent chromosome will be transferred to the next step unchanged. The displacement mutation technique is used to execute the mutation operator. Two genes on the parent chromosome are randomly selected for fulfilling this function. Followed by selecting the two desired genes, their numerical values will be transferred to each other.

3.2.1. Replacement and stopping criteria

The generational update is used in the designed genetic algorithm. In fact, the offspring produced at each stage of the genetic algorithm will replace their parents in the next population. In other words, the offspring is generated as many as the number of chromosomes in the current population and it makes up the next population in the genetic algorithm. In addition, the number of iterations of the genetic algorithm is considered as a stopping criterion, which is represented by the variable iteration.

Ensemble classifier pseudocode

In this section, ECGA was designed based on the expert systems of the decision tree, random forest, K-nearest neighbor, and naive Bayes. Algorithm 1 shows the basic genetic ensemble classifier pseudocode. This classifier is used to diagnose diabetes.

3.3. Environmental conditions

In this section, the impact of environmental conditions on diabetes is examined. In fact, it investigates how each training dataset operates on the test dataset. Let $TR = (tr_1, tr_2, \cdots, tr_s)^t$ be a set of trains with different samples but similar features, and $TE = (te_1, te_2, \cdots, te_s)^t$ a set of tests from the corresponding trains so that te_s. Here, s represents the number of regions. In case Alg is a classifier algorithm such as Decision Tree, Nave Bayesian, etc., Eq. (4) is written as:

$$L_{ij} = Alg(tr_i, te_j), \qquad i, j = 1, \cdots, s, \tag{4}$$

where L_{ij} is the output labels by the classifier algorithm Alg. If we assume that Lte_j, $j = 1, \cdots, s$, can be the real labels of the dataset te_j, the output quality size $(L_{ij})Alg$ is shown in Eq. (5):

$$R_{ij} = Accuracy(L_{ij}, Lte_j). \tag{5}$$

It should be proved that if $R_{ii} > R_{ij}$, $i \neq j$, then te_i is better suited for tr_i, indicating that the conditions

01. Inputs: Size, P_c, P_m and Iteration.
02. Generate initial population, randomly.
03. for $i = 1$ to iteration
04. Evaluate fitness value of each chrommsome based on equation (2).
05 Select number of chrommsomes with roulette wheel strategy.
06. Generate ran_c. \\ $ran_c$is a random number between 0 and 1.
07. if $ran_c <= P_c$
08. Operate two-point crosover on selected choromosomes.
09. end if
10. Generate ran_m. \\ $ran_m$is a random number between 0 and 1.
11. if $ran_m <= P_m$
12. Operate swap mutaition on crosover choromosomes.
13. end if
14. end for
15. Determine the best choromosome in the population.
16. output: The best coefficients for expert systems.

Algorithm 1. ECGA pseudocode.

01. Inputs: A_i, $i \longleftarrow 1$ to n. \\ A_i represents the area i, n represents the number of areas s.
02. Generate TrainData and TestData for each area (A_i).
03. Set Tr_i for $Train(A_i)$ and Te_i for $Test$ (A_i).
04. for $i \longleftarrow 1$ to n
05. for $j \longleftarrow 1$ to n
06. $P_{ij} \longleftarrow Algorithm\ (Te_i, Tr_j)$ \\ P_{ij} is accuracy of classifier on (Te_i, Tr_j)
07. end for
08. end for
09. if $P_(ij)(i = j) >= P_(ij)(i \neq j)$
10. This means that environmental conditions are influential.
11. end if
12. Output: accuracy of (Te_i, Tr_j)

Algorithm 2. The pseudocode to assess the impact of environmental conditions.

in the area are effective concerning the disease. The general structure of investigating this case is typically in the form of the pseudo-code presented in Algorithm 2.

Three different models were designed to investigate the effect of environmental conditions (by examining three different areas) on diabetes. The first, second, and third models were built based on the $Train_1$, $Train_2$, and $Train_3$ datasets, respectively. Figure 6 shows the solution for examining the effects of environmental conditions on diabetes in three different areas. In this examination, three steps should be taken into account:

Step 1: The three designed models are evaluated on the $Test_1$ dataset, and measurement of diagnostic accuracy is generated as the output. In case $Output_{11}$ outperforms $Output_{21}$ and $Output_{31}$, it can be concluded that the environmental conditions affect diabetes;

Step 2: The three designed models are evaluated on the dataset $Test_2$, and measurement of the diagnostic accuracy is generated as output. If $Output_{22}$ outperforms $Output_{12}$ and $Output_{32}$, it can be concluded that the environmental conditions affect diabetes;

Step 3: The three designed models are evaluated on the dataset $Test_3$, and measurement of diagnosis accuracy is generated as the output. In case $Output_{33}$ outperforms $Output_{13}$ and $Output_{23}$, it can be concluded that environmental conditions affect diabetes.

4. Experimental results

4.1. Datasets

Table 2 presents some information about the dataset.

Table 2. The information related to the dataset.

Dataset	Total samples	Training samples	Testing samples
DB_1	500	400	100
DB_2	500	400	100
DB_3	500	400	100
DB	1500	1200	300

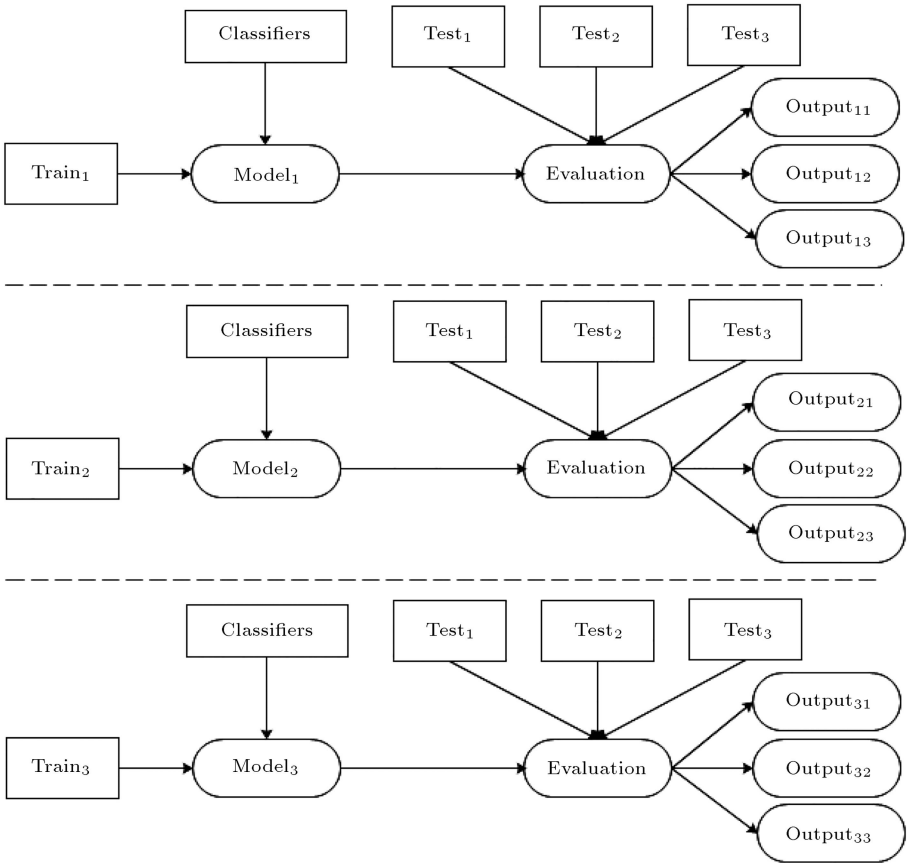

Figure 6. Investigation of the effect of the environmental conditions on diabetes.

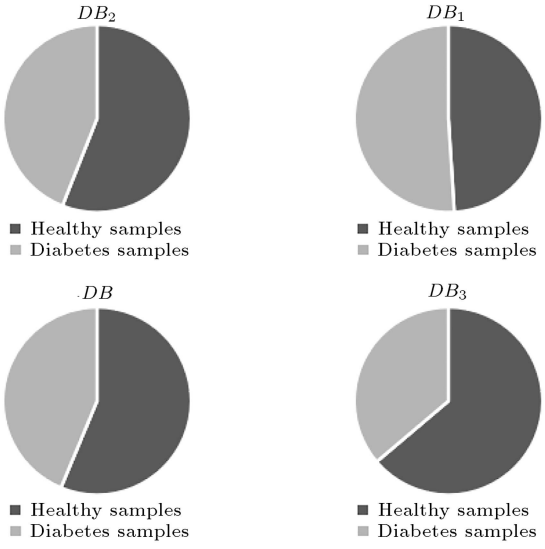

Figure 7. Frequency distribution of samples in the dataset.

It shows the total number of samples in the dataset. Generally, 80% and 20% of the samples in each dataset are used for training and testing datasets, respectively. In addition, Figure 7 shows the frequency distribution of the samples in the dataset. Obviously, each sample in the dataset has either a patient or healthy class.

Table 3. Initial values of the parameters of the genetic algorithm.

No.	Parameter	Value
1	Chromosome length	10
2	Population size	100
3	Imax	300
4	P_c	0.8
5	P_m	0.01

4.2. Initialization parameters

Population size is considered constant in the genetic algorithm. Table 3 shows the initial values of the parameters of the genetic algorithm.

4.3. Result

Table 4 shows the diagnostic accuracy of different classifiers applied to the DB dataset. As shown in Table 4, the ECGA has an accuracy rate of 87.91%, indicating that ECGA outperformed ten expert systems. It also improved the diagnostic accuracy criterion by 1.64%, compared to the best expert system. Figure 8 shows the performance of the classifiers in the DB dataset in terms of the error rate criterion. ECGA had the lowest error rate among the expert systems computed as 12.09%.

Table 4. The classifiers' diagnostic accuracy.

No.	Expert system	Abbreviation	Accuracy (%)
1	Decision tree	DT	82.04
2	Random forest	RF_1	85.20
3	Random forest	RF_2	85.09
4	Random forest	RF_3	85.20
5	Random forest	RF_4	85
6	Nave Bayesian	NB	85.40
7	K-nearest neighbors	KNN_1	81.69
8	K-nearest neighbors	KNN_2	84.60
9	K-nearest neighbors	KNN_3	85.31
10	K-nearest neighbors	KNN_4	86.27
11	Ensemble classifier based on genetic algorithm	$ECGA$	87.91

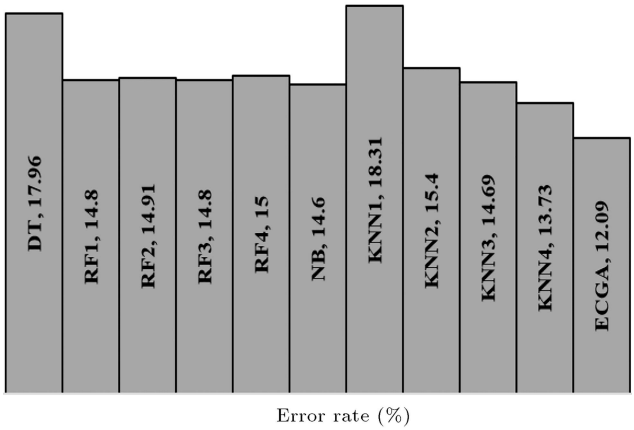

Figure 8. Performance of classifiers in terms of error rate criterion.

Table 5. The results obtained from evaluation of the first step.

Expert system	$Output_{11}$	$Output_{21}$	$Output_{31}$
DT	**94.54**	88.21	90.37
RF_1	**96.63**	92.52	90.23
RF_2	**96.63**	92.72	90.37
RF_3	**96.70**	92.79	90.64
RF_4	**96.56**	92.59	90.57
NB	**94.34**	88.21	90.37
KNN_1	**93.73**	86.06	87.34
KNN_2	**94.61**	89.76	91.24
KNN_3	**95.01**	91.31	91.98
KNN_4	**95.69**	92.45	92.12
$ECGA$	**96.97**	92.94	94.07

Table 5 shows the results from the first step of examining the impact of the environmental conditions on diabetes. As observed, $Output_{11}$ outperformed $Output_{21}$ and $Output_{31}$ in all expert systems. Therefore, based on the assessment made in the first step, it can be concluded that environmental conditions affect diabetes.

Table 6. Results of evaluating the second step.

Expert system	$Output_{12}$	$Output_{22}$	$Output_{32}$
DT	79.66	**82.96**	74.94
RF_1	81.81	**84.31**	76.49
RF_2	81.75	**84.51**	77.03
RF_3	81.75	**84.57**	77.37
RF_4	81.61	**84.64**	77.10
NB	80.80	**81.21**	78.95
KNN_1	78.45	**78.99**	77.17
KNN_2	81.21	**81.27**	78.04
KNN_3	82.22	**82.96**	80.40
KNN_4	82.76	**84.51**	81.34
$ECGA$	86.26	**86.66**	82.66

Table 7. Results of evaluating the third step.

Expert system	$Output_{13}$	$Output_{23}$	$Output_{33}$
DT	82.89	81.01	**86.26**
RF_1	83.97	83.70	**88.14**
RF_2	84.10	83.63	**89.36**
RF_3	84.10	84.17	**88.48**
RF_4	83.83	83.50	**88.62**
NB	83.97	79.73	**86.66**
KNN_1	83.16	79.19	**85.05**
KNN_2	82.76	80.94	**85.11**
KNN_3	83.30	82.15	**86.53**
KNN_4	84.24	83.23	**86.80**
$ECGA$	86.53	86.01	**91.26**

Table 6 shows the results of the second step of the environmental assessment and its relation to diabetes. As observed, $Output_{22}$ outperforms $Output_{12}$ and $Output_{32}$ in all expert systems. Therefore, based on the assessment made in the second step, it can be concluded that the environmental conditions have significant impacts on diabetes.

Table 7 shows the results of the third step of examining the environmental conditions impact on

Table 8. The performance of expert systems in evaluating the environmental conditions impact on diabetes.

Expert system	$Output_{11}$	$Output_{21}$	$Output_{31}$	$Output_{12}$	$Output_{22}$	$Output_{32}$	$Output_{13}$	$Output_{23}$	$Output_{33}$
DT	94.54	88.21	90.37	79.66	82.96	74.94	82.89	81.01	86.26
RF_1	96.63	92.52	90.23	81.81	84.31	76.49	83.97	8.70	88.14
RF_2	96.63	92.72	90.37	81.75	84.51	77.03	84.10	83.63	89.36
RF_3	96.70	92.79	90.64	81.75	84.57	77.37	84.10	84.17	88.48
RF_4	96.56	92.59	90.57	81.61	84.64	77.10	83.83	83.50	88.62
NB	94.34	88.21	90.37	80.80	81.21	78.95	83.97	79.73	86.66
KNN_1	93.73	86.06	87.34	78.45	78.99	77.17	83.16	79.19	85.05
KNN_2	94.61	89.76	91.24	81.21	81.27	78.04	82.76	80.94	85.11
KNN_3	95.01	91.31	91.98	82.22	82.96	80.40	83.30	82.15	86.53
KNN_4	95.69	92.45	92.12	82.76	84.51	81.34	84.24	83.23	86.80
$ECGA$	**96.97**	**94.94**	**94.07**	**86.26**	**86.66**	**82.66**	**86.53**	**86.01**	**91.26**

diabetes according to which $Output_{33}$ outperforms $Output_{13}$ and $Output_{23}$ in all expert systems; hence, the environmental conditions affect diabetes.

Table 8 makes a comparison of the performance of the expert systems to assess the effect of environmental conditions on diabetes. According to this table, ECGA had the highest degree of diagnostic accuracy among all nine outputs of the expert systems. For this reason, ECGA exhibits the best performance among all expert systems.

5. Conclusion

Diabetes is considered as one of the serious diseases threatening the global health. In recent decades, it has spread rapidly and caused serious damage to vital systems of human body, especially the nervous system. For this reason, the application of medical data mining for diagnosing and analyzing diabetes has gained significance. The present study aimed to investigate the impacts of the environmental conditions on diabetes based on Ensemble Classifier based on Genetic Algorithm (ECGA). Findings revealed that ECGA exhibited the best performance with the diagnostic accuracy of 87.91% in the diabetes dataset, compared to the expert systems of the decision tree, random forest, K-nearest neighbor, and naive Bayes. In addition, the results confirmed the considerable impacts of environmental conditions on diabetes.

References

1. Bashir, S., Qamar, U., and Khan, F.H. "Intellihealth: a medical decision support application using a novel weighted multi-layer classifier ensemble framework", *J. Biomed. Inform. X*, **59**, pp. 185–200 (2016).
2. Carter, J.A., Long, C.S., Smith, B.P., et al. "Combining elemental analysis of toenails and machine learning techniques as a non-invasive diagnostic tool for the robust classification of type-2 diabetes", *Expert Syst. Appl.*, **115**, pp. 245–255 (2019).
3. Atlas, D. "International diabetes federation", *idf Diabetes Atlas*, 1st Edn., Brussels, Belgium: International Diabetes Federation (2000).
4. Atlas, D. "International diabetes federation", *idf Diabetes Atlas*, 2nd Edn., Brussels, Belgium: International Diabetes Federation (2003).
5. Atlas, D. "International diabetes federation", *idf Diabetes Atlas*, 3rd Edn., Brussels, Belgium: International Diabetes Federation (2006).
6. Atlas, D. "International diabetes federation", *idf Diabetes Atlas*, 4th Edn., Brussels, Belgium: International Diabetes Federation (2009).
7. Atlas, D. "International diabetes federation", *idf Diabetes Atlas*, 5th Edn., Brussels, Belgium: International Diabetes Federation (2011).
8. Atlas, D. "International diabetes federation", *idf Diabetes Atlas*, 6th Edn., Brussels, Belgium: International Diabetes Federation (2013).
9. Atlas, D. "International diabetes federation", *idf Diabetes Atlas*, 7th Edn., Brussels, Belgium: International Diabetes Federation (2015).
10. Atlas, D. "International diabetes federation", *idf Diabetes Atlas*, 8th Edn., Brussels, Belgium: International Diabetes Federation (2017).
11. Zilbermint, M. "Diabetes and climate change", *J Community Hosp Intern Med Perspect*, **10**(5), pp. 409–412 (2020).

12. Huang, Y., McCullagh, P., Black, N., et al. "Feature selection and classification model construction on type 2 diabetic patients data", *Artif Intell Med.*, **41**(3), pp. 251–262 (2007).

13. Kahramanli, H. and Allahverdi, N. "Design of a hybrid system for the diabetes and heart diseases", *Expert Syst. Appl.*, **35**(1–2), pp. 82–89 (2008).

14. Temurtas, H., Yumusak, N., and Temurtas, F. "A comparative study on diabetes disease diagnosis using neural networks", *Expert Syst. Appl.*, **36**(4), pp. 8610–8615 (2009).

15. Patil, B.M., Joshi, R.C., and Toshniwal, D. "Hybrid prediction model for type-2 diabetic patients", *Expert Syst. Appl.*, **37**(12), pp. 8102–8108 (2010).

16. Ganji, M.F. and Abadeh, M.S. "A fuzzy classification system based on antcolony optimization for diabetes disease diagnosis", *Expert Syst. Appl.*, **38**(12), pp. 14650–14659 (2011).

17. Çalişir, D. and Doğantekin, E. "An automatic diabetes diagnosis system based on lda-wavelet support vector machine classifier", *Expert. Syst. Appl.*, **38**(7), pp. 8311–8315 (2011).

18. Aslam, M.W., Zhu, Z., and Nandi, A.K. "Feature generation using genetic programming with comparative partner selection for diabetes classification", *Expert Syst. Appl.*, **40**(13), pp. 5402–5412 (2013).

19. Seera, M. and Lim, C.P. "A hybrid intelligent system for medical data classification", *Expert Syst. Appl.*, **41**(5), pp. 2239–2249 (2014).

20. Gorzałczany, M.B. and Rudzinski, F. "Interpretable and accurate medical data classification–a multiobjective genetic-fuzzy optimization approach", *Expert Syst. Appl.*, **71**, pp. 26–39 (2017).

21. Kavakiotis, I., Tsave, O., Salifoglou, A., et al. "Machine learning and data miningmethods in diabetes research", *Comput Struct. Biotechnol. J.*, **15**, pp. 104–116 (2017).

22. Kumari, S., Kumar, D., and Mittal, M. "An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier", *International Journal of Cognitive Computing in Engineering*, **2**, pp. 40–46 (2021).

23. Zheng, T., Xie, W., Xu, L., et al. "A machine learning-based framework to identify type 2 diabetes through electronic health records", *Int. J. Med. Inform.*, **97**, pp. 120–127 (2017).

24. Gupta, D., Choudhury, A., Gupta, U., et al. "Computational approach to clinical diagnosis of diabetes disease: a comparative study", *Multimed Tools Appl.*, **80**, p. 126 (2021).

25. Rahman, M., Islam, D., Mukti, R.J., et al. "A deep learning approach based on convolutional lstm for detecting diabetes", *Comput. Biol. Chem.*, pp. 107–329 (2020).

26. Pranto, B., Mehnaz, S., Mahid, E.B., et al. "Evaluating machine learning methods for predicting diabetes among female patients in Bangladesh", *Information*, **11**(8), p. 374 (2020).

27. Kannadasan, K., Edla, D.R., and Kuppili, V. "Type 2 diabetes data classification using stacked autoencoders in deep neural networks", *Clin Epidemiol Glob Health*, **7**(4), pp. 530–535 (2019).

28. Patra, R. and Bonomali, K. "Analysis and prediction of pima Indian diabetes dataset using sdknn classifier technique", In *IOP Conference Series: Materials Science and Engineering*, IOP Publishing, **1070**, p. 012059 (2021).

29. Khanam, J.J. and Foo, S.Y. "A comparison of machine learning algorithms for diabetes prediction", *ICT Express*, **7**(4), pp. 432–439 (2021).

30. Kaul, S. and Kumar Y. "Artificial intelligence-based learning techniques for diabetes prediction: Challenges and systematic review", *SN Comput. Sci.*, **1**(6), pp. 1–7 (2020).

31. Sangien, T., Bhat, T., and Khan, M.S. "Diabetes disease prediction using classification algorithms", *Internet of Things and Its Applications*, Springer, pp. 185–197 (2022).

Biographies

Fateme Khademi received her BSc degree in Computer Engineering from Islamic Azad University, Sari Branch, Iran in 2008. She received her MSc degree in Computer Science from the same university in 2012. She is currently a PhD student at Islamic Azad University, Sari Branch. Her research interests include data mining algorithms and applications as well as optimization algorithms.

Homayun Motameni received his BSc degree in Computer Engineering-Software engineering from Shahid Beheshti University and MSc degree in Computer Engineering and Machine Intelligence from Islamic Azad University and Science and Research Branch in 1995 and 1998, respectively. He received PhD degree in Computer Engineering (Software Engineering) from Islamic Azad University-Science and Research Branch in 2007. His Current research interests include evolution algorithms, Petri Net, and software systems modeling and evaluation using Petri Net, and machine learning.

Mohsen Rabbani He was born in 1967 in Iran. He received his BSc degree in Applied Mathematics (Operations Research) from Tehran University in 1991, MSc degree in Applied Mathematics (Numerical Analysis) from Iran University of Science and Technology

(I.U.S.T) in 1998, and PhD degree in Applied Mathematics (Integral Equations) from I.U.S.T in 2008. He is currently an Associate Professor at the Department of Applied Mathematics, Azad university of Iran. His favorite research subjects are integral equations, data mining, wavelet, numerical solution of ordinary, and partial differential equations.

Ebrahim Akbari received his PhD degree in Computer Science from Universiti Teknologi, Malaysia in 2015. He is now a Professor in Computing Science at the Department of Computer Engineering, Islamic Azad University, Sari Branch. His research interests are data analysis, data mining algorithms and applications, machine learning, and pattern recognition.