www.ijsrnsc.org

Available online at www.ijsrnsc.org

IJSRNSC

Volume-9, Issue-3, June 2021
Research Paper

Int. J. Sc. Res. in Network Security and Communication

E-ISSN:2321-3256

Diagnosis of Diabetes Using Bee Colony Algorithm and Fuzzy Decision Tree

M. Mojarad[1*], E. Hajizadegan[2], M. Gurkani[2]

[1]Dept. of Computer Engineering, Firoozabad Branch, Islamic Azad University, Firoozabad, Iran
[2]Dept. of Computer Engineering, Liyan Institute of Education, Bushehr, Iran

[*]*Corresponding Author: m.mojarad@iauf.ac.ir, Tel.: +98-91788-63397*

Received: 30/Mar/2021, Accepted: 20/Jun/2021, Published: 30/Jun/2021

***Abstract*—** Today, in medical knowledge, collecting a lot of data about different diseases is very important. Medical centers collect this data for various purposes. One of the goals of using this data is to research these data and obtain useful results and patterns in relation to diseases. The large volume of this data and the confusion used to overcome this problem to obtain useful relationships between risk factors in diseases. In this study, due to the importance of diabetes in medicine, the aim is to present a hybrid model using fuzzy decision tree and bee cloning algorithm to increase the accuracy of diagnosis. The proposed method is called ABC-FDT. In ABC-FDT, the number of optimal fuzzy sets for each feature is considered so that the best segmentation for the features is provided with the two goals of accuracy and reduction of complexity. PID diabetes data set and classification methods based on ID3, C4.5 and CART rules were used for evaluation. The results indicate the superiority of ABC-FDT algorithm in terms of rules of number, accuracy, sensitivity and specificity. ABC-FDT also outperformed the recently introduced GAANN, QFAM-GA, and SM-RuleMiner algorithms by 10.1, 5.4, and 7.8 percent, respectively.

***Keywords*—**Diagnosis of diabetes, bee colony algorithm, fuzzy decision tree.

I. INTRODUCTION

Diabetes is one of the most common diseases in the world. One of the main problems related to this disease is its lack of timely and correct diagnosis. The incidence of diabetes has doubled worldwide in the last ten years. About 200 million people are affected by the disease, and the prevalence of diabetes in the world is increasing by about six percent annually. More than two million people in Iran are infected with this disease [1]. In this study, we investigate the relationship between complications observed in diabetic patients and some of their characteristics such as blood sugar, blood pressure, age and family history of patients. The aim of this study was to predict patients' complications based on the symptoms observed in them [2]. The main problem with diabetes is the lack of timely diagnosis or general weakness in the diagnosis of this disease, which is also due to the lack of proper selection of the model by the doctor or the lack of proper use of standard models [3]. Therefore, providing a method that can help any person in the correct diagnosis of the disease or not can be an important step in the prevention and control of this disease, especially in its early stages. Data mining is a way to automatically analyze data and identify hidden patterns that are not possible manually. Data mining can be used effectively to predict and diagnose diseases quickly and cheaply. The importance of predicting diabetes is that once the patient is aware of it, they can prevent its devastating effects by modifying their diet and exercise.

Therefore, in this study, data mining is used to analyze, diagnose and predict diabetes. In this way, by using data mining algorithms and methods along with different patients' information, a way to analyze, diagnose and predict different diseases is provided. In fact, raw data is patient information and methods used in data mining sciences that lead to the production of useful knowledge for medical sciences. Studying these articles can be helpful and useful for people who work in the field of medical and health data mining [4]. According to some, medical science is a science based on statistics, and many of the solutions offered to patients are obtained through statistical analysis methods and methods. This claim and related facts are examined in this series of articles, which will be produced under the title of using data mining to analyze diseases in different cases.

So far, various intelligent methods have been proposed to solve this fundamental problem around the world, including the use of evolutionary methods, fuzzy algorithms, pattern recognition in feature extraction, and neural networks [5]. Therefore, one of the problems that diabetic patients currently face is the weakness in the diagnosis of this disease in its early stages. Therefore, in this study, we have tried to use a combination of bee colony algorithms and decision tree. Provide a solution to identify patients. In the proposed hybrid method, first the Artificial Bee Colony Algorithm (ABC) is used to select the best effective features in the diagnosis of diabetes. In the next step, using an innovative isolation technique, an intelligent

classification model based on a fuzzy decision tree is presented.

In the continuation of this research, we will examine some of the works done in section II. In Section III, the proposed method based on bee colony algorithm and fuzzy decision tree for diabetes is presented. The results of the evaluation of the proposed method and its discussion are given in Section IV, and finally, Section V conclusion and future work is given.

II. RELATED WORKS

Jane et al. (2015) proposed the improvement of diabetes prediction using the intelligent fuzzy system [6]. Fuzzy reasoning has been used to improve the prognosis of diabetes, which diagnoses diabetes based on patients' knowledge and experience. Raw data is converted to fuzzy data, then using the IF-THEN rule model, the fuzzy input is converted to fuzzy output.

In [7], a genetic algorithm is used to predict diabetes. Fuzzy systems are used to solve a wide range of problems in the range of different applications of genetic algorithms for design. The proposed fuzzy system enables the introduction of learning and adaptation capabilities. Neural networks are used effectively to learn membership functions. Diabetes occurs worldwide, but type 2 is more common in most developed countries. There is a further increase in prevalence. However, many patients in Asia and Africa are expected to be discovered by 2030.

In [8], a Genetic Programming (GP) method for the diagnosis of diabetes is proposed. Genetic programming uses evolutionary computing and discipline computational programs to make decisions. In this system, genetic programming is considered for the evaluation and classification of diabetic patients, which is done based on previous knowledge entered in this system. In addition to data mining, genetic programming is used to categorize diabetic, non-diabetic, and pre-diabetic patients. This system provides a multidisciplinary classification for diabetes and acts as the doctor's second theory. This system alerts patients and the doctor can take the necessary steps. So it saves time. With the growth of the fuzzy system and logical behavior, this multi-group GP method can coordinate physicians, especially those with little experience.

In [9], modified AIRS2 (Artificial Immune Recognition System 2) is used where the adjacent point algorithm is replaced with an adjacent fuzzy point K to improve the accuracy of diabetes diagnosis. The diabetes database used in the simulation is taken from the UCI learning repository. The performance of AIRS2 and MAIRS2 (Modified AIRS2) is evaluated according to classification accuracy, sensitivity values and specialization. The highest classification accuracy is obtained when AIRS2 and AIRS2 are used with fuzzy KNN algorithm using decimal evaluation. The accuracy obtained is 82.69% and 89.10%.

In [10], a fuzzy classification for diabetes is designed using an optimized artificial bee colony algorithm. In this research, a new bee algorithm is proposed by adding a mutation operator to ABC to improve its performance. When the current best solution for genetic diversity is not improved, a hybrid operator called BLX-α uses a genetic algorithm to increase ABC diversity. In order to evaluate the diabetes dataset, the UCI machine learning repository was used, which is very promising in terms of speed, sensitivity and features compared to previous methods.

The GAANN (Genetic Algorithm Artificial Neural Network) method has been proposed by Ahmad et al. (2015) for the diagnosis of cancer. In this study, we present the results of this model on the diabetes database for comparison [11]. GAANN is a classification model based on genetic algorithm and neural network ANN, which uses a genetic algorithm to select features and simultaneously optimize neural network weights. This algorithm uses three techniques GAANN_RP, GAANN_LM and GAANN_GD to adjust ANN weights and shows the better performance results of GAANN_RP.

The QFAM-GA (Q-Learning Fuzzy ARTMAP- Genetic Algorithm) method was proposed by Pourpanah et al. (2016) to diagnose diabetes [12]. QFAM-GA is a two-step hybrid model that uses fuzzy methods and genetic algorithms to classify data and extract rules. In the first stage, it uses ARTMAP fuzzy classification (FAM) with Q-learning for incremental learning and creating a classification model. This step is known as QFAM for short. In the second stage, the genetic algorithm uses the QFAM model to derive the law. A hybrid model called QFAM-GA can use fuzzy rules to predict the target class of input data samples. The study also uses Q-values to reduce the number of samples produced by QFAM to reduce network complexity.

The SM-RuleMiner method was proposed by Cheruku et al. (2017) to diagnose diabetes [13]. Law-based classification systems have been widely used to diagnose diabetes, but the challenge of these methods is to produce a set of desirable rules overcoming accuracy, sensitivity, and specificity. To solve this problem, in this study, the SM-RuleMiner model, which is a spider monkey optimization algorithm based on Miner's law, was proposed to classify diabetes. In this model, a new objective function is also developed to generate an optimal set of rules by balancing the criteria of accuracy, sensitivity and specificity.

III. THE PROPOSED METHOD

In the classical fuzzy decision tree, the values of each feature are fuzzy according to a membership degree function and then created based on the segmentation of the rules database. In each rule, according to the characteristics of the input sample and examining them in its preamble, it identifies the output class of the sample, which is "sick" and "not sick".

Membership degree functions can include one or more fuzzy sets with different decision values. For example, a two-state membership degree function has two fuzzy sets, Low and High. The more cases (the number of fuzzy subsets) there are, the more situations the decision process will have, thus increasing the accuracy in solving the problem. But this increase leads to complexity and increases the parameters of the problem.

Fuzzy decision tree algorithms classify feature values for each feature according to a fuzzy membership degree function, and consider a fuzzy set for each category that has a numerical range. Therefore, the more of these categories, the more samples are properly covered and thus the accuracy increases, but this increase leads to the creation of very large trees that for some large data sets are not possible.

Our goal in this study was to consider the optimal number of fuzzy sets for each property. In fact, we want to assign different membership degree functions to each feature so that different states of the feature values are covered. Therefore, two goals are pursued in this issue; the first goal is to increase the classification accuracy and the second goal is to reduce the number of fuzzy sets for each feature. For example, for the first feature, a membership degree function with three values A1, A2 and A3 is assigned, and for the second feature, a membership degree function with two values, B1 and B2, is assigned. Therefore, we assign a membership class function with the number of a values to each feature according to the separation values and its importance in sampling the samples to optimize the two specified objectives.
Another idea of this research is to select features related to classification. That is, based on the information gain calculated in the decision tree, we will try to reduce the branches and the depth of the tree, and as a result, the number of features. In order to assign different degree degree functions, we use the ABC optimization algorithm.
In the proposed method, we consider the diabetes data set as input. Then, based on 10-Fold validation, we divide the data set into two training (E^T) and experimental (E^P) datasets. We use E^T to create fuzzy rules models and databases to classify data, and E^P to evaluate and test fuzzy rules classification models. In the next step, we use the bee colony algorithm (ABC) to determine the optimal fuzzy sets for each feature. The ABC algorithm tries to find the best segmentation for features with the two goals of accuracy and reduction of complexity by creating several fuzzy sets in the form of three worker bees (EB), observer (OB) and explorer (SB) and performing various searches. Therefore, the validity of the solutions generated by a two-objective function is calculated on the E^T dataset. To calculate the suitability, the E^T data set is fuzzy according to the fuzzy set of each solution.

In the ABC algorithm, to improve the search process, it assigns each solution produced by the worker bee with the Limit threshold if the competency is not improved. This process is repeated with one termination condition until the maximum number of iterations is reached and the solutions are optimized. At each step, we store the best fuzzy set (solution) found in the BestGlobal variable and use this solution to create a database of fuzzy rules. Finally, the simulation results are reported by applying the experimental data set (E^T) and evaluating the generated rules.
The flowchart of the proposed method, called the ABC-FDT method, is shown in Figure 1.

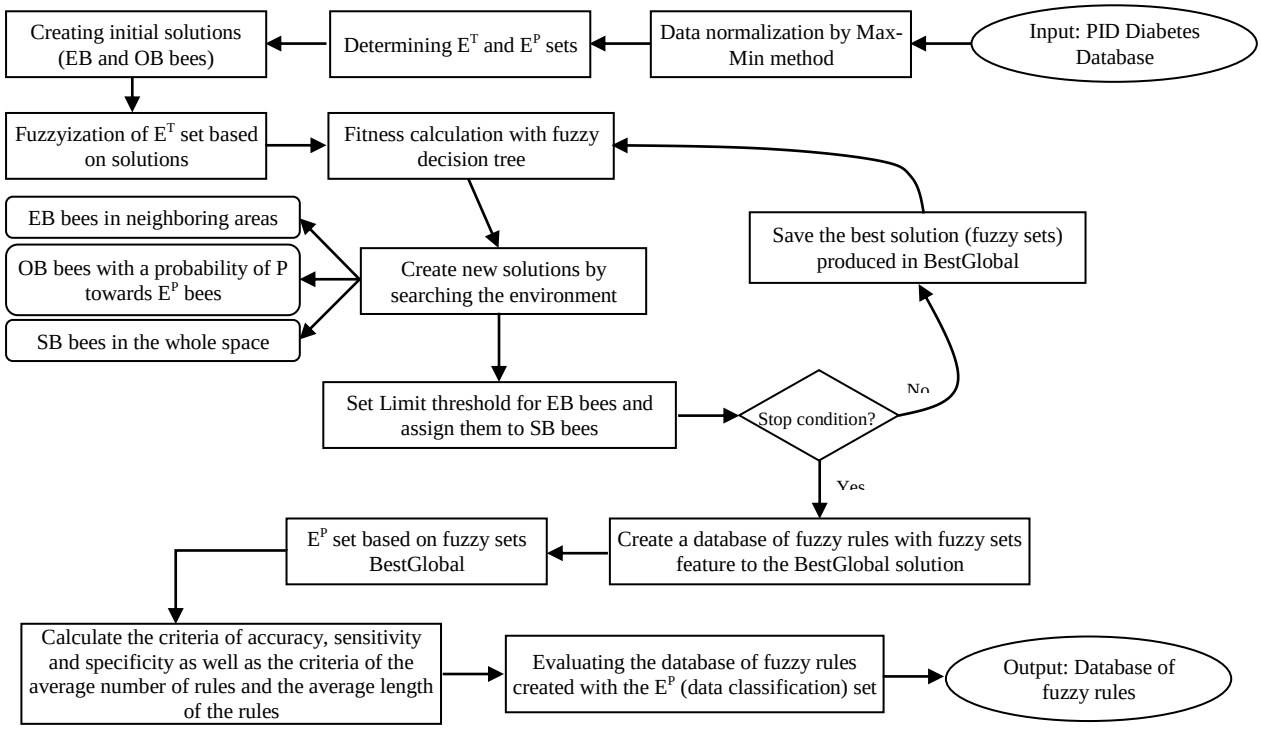

Figure 1. Flowchart of the proposed method

Given that in the present problem we are looking for a number of fuzzy subsets for each property, we use a vector of the length of the properties to display the answers. Figure 2 shows an example of the structure of the answer in the problem.

U_1	U_2	. . .	U_i	...	U_M

Figure 2. The structure of representation solutions

In the proposed structure, each solution for the M property is equivalent to a vector of length M, and U_i represents the number of fuzzy subsets for the i-th property. In this study, the number of N_P solutions is randomly generated, half of which play the role of worker bees and the other half the role of observer bees.

If $x_{i,j}$ represents the i-th solution ($i = 1,2, \dots, NP$) in the j-th dimension of the problem ($j = 1,2, \dots, M$), during the evolution cycle of the ABC algorithm Three movements in space are performed by worker, observer and discoverer bees. The following describes the changes made to the content of the solutions and the creation of new solutions.

1. Moving worker bees to neighboring areas to produce new solutions: All worker bees produce a new solution based on Eq. (1).

$$x_{i,j} = x_{i,j} + \varphi_{i,j}\left(x_{i,j} - x_{k,j}\right), \quad \forall j = 1,2, \dots, M \tag{1}$$

Where, $\varphi_{i,j}$ is a random number with a uniform distribution in the interval $[-1, +1]$ that controls the production of the position of the neighboring solution around $x_{i,j}$. j is the symbol of the solution dimensions and k is the solution index which is randomly selected from the colony.

2. The movement of observer bees with a probability of P towards worker bees: An observer bee chooses one of the solutions created by worker bees with a probability of P and creates a new solution towards the solution of the worker bee. The process of producing a new solution is similar to that of a worker bee, with only the k-index being determined as Eq. (2).

$$P_k = \frac{fitness_k}{\sum_{k=1}^{NP_{EB}} fitness_k}, \quad \forall\, k \in EB \tag{2}$$

Where, P_k is the probability of being selected by a control bee. $fitness_k$ indicates the degree of competence of the k-worker bee and NP_{EB} shows the number of worker bees in the colony.

3. Random movement of detective bees throughout the search space: After all observer bees in neighboring locations have generated movement and new solutions, the generated solutions are examined to see if they should be abandoned. If the number of cycles that a worker bee cannot improve is greater than the predetermined limit, the worker bee solution is considered as an unimproved solution. The worker bee related to this solution becomes an explorer bee and creates a random search in the problem space. Eq. (3) shows the random search of the discoverer bee.

$$x_{i,j} = x_{j-min} + rand \times \left(x_{j-min} - x_{j-min}\right), \quad \forall j = 1,2, \dots, M \tag{3}$$

Where, x_{j-min} and x_{j-min} are the minimum and maximum values allowed for the j-th after the problem, respectively.

IV. RESULTS AND DISCUSSION

Extensive experiments have been performed in this section to investigate the superiority of the proposed method. In this paper, MATLAB software version 2019a has been used to simulation. Simulation and all tests were performed using a 2.4 GHz Intel Cor i7 CPU and 8 GB of RAM. The simulation results of the proposed method called "ABC-FDT" are shown in all experiments. In this paper, GAANN, QFAM-GA and SM-RuleMiner algorithms are used to compare and evaluate the performance of the proposed algorithm. Also, real PID diabetes databases are used for simulation and comparison, which is from the UCI repository. The PID data set contains 8 features and 768 samples.

The proposed method tries to create a fuzzy decision tree using the best bee cloning algorithm with the best segmentation in terms of the values of each feature. The division of each property is done according to the trapezoidal membership function with different threshold values. Table I shows the final segmentation results for 8 properties in 10-fold.

Table 1. Segmentation results for 8 features

Feature number	Number of sections	Segmentation symbol	Segmentation threshold values
1	9	A	{0, 0.11, 0.22, 0.33, 0.44, 0.55, 0.66, 0.77, 0.88, 1}
2	7	B	{0, 0.143, 0.286, 0.429, 0.572, 0.715, 0.858, 1}
3	7	C	{0, 0.143, 0.286, 0.429, 0.572, 0.715, 0.858, 1}
4	5	D	{0, 0.2, 0.4, 0.6, 0.8, 1}
5	5	E	{0, 0.2, 0.4, 0.6, 0.8, 1}
6	6	F	{0, 0.17, 0.34, 0.5, 0.68, 0.84, 1}
7	6	G	{0, 0.17, 0.34, 0.5, 0.68, 0.84, 1}
8	5	H	{0, 0.2, 0.4, 0.6, 0.8, 1}

Based on the classifications made, Figure 3 shows a portion of the final fuzzy decision tree for modeling the Diabetes dataset for 8 features. This tree is presented according to the 10-fold output result.

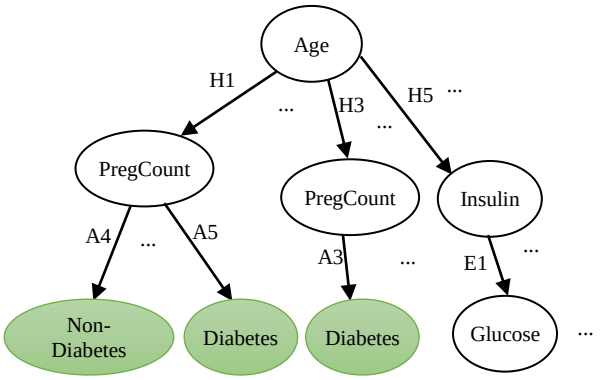

Figure 3. A small part of the final fuzzy decision tree

In order to evaluate the performance of the proposed method, a comparison has been made with the three recently introduced GAANN, QFAM-GA and SM-RuleMiner methods. The results are shown in Table II according to the 10-Fold average for the criteria of average number of rules, average length of rules as well as three criteria of accuracy, sensitivity and specificity.

The results show that the proposed ABC-FDT method has the best prediction accuracy with 96.87% compared to other methods. Also, the best sensitivity of 95.12% and the best specificity criterion of 100.0% have been reported compared to GAANN and SM-RuleMiner methods. Although the proposed method has less speed than the two methods QFAM-GA and SM-RuleMiner by increasing the number of rules, but in general, considering all the cases, it has been able to achieve better accuracy.

Table 2. Comparing the performance of the proposed method with other methods

Algorithms	Average Number of Rules	Average Length of Rules	Accuracy (%)	Sensitivity (%)	Specificity (%)
GAANN [11]	-	-	88.02	86.67	88.64
QFAM-GA [12]	6.0	5.4	91.91	-	-
SM-RuleMiner [13]	4.1	2.02	89.87	94.6	80.11
ABC-FDT	468.1	4.95	96.87	95.12	100.0

V. CONCLUSIONS AND SUGGESTIONS

In Medical data mining can be a great help to the medical community in discovering the hidden patterns of meaningful and important relationships between the large amount of data and information available according to the severity of the patient population and ultimately increase the accuracy of disease prediction; Help simplify the treatment process and reduce the costs associated with it. In this study, a hybrid approach based on bee cloning algorithm as well as fuzzy decision tree for diagnosing diabetes was presented. In this problem, two goals were pursued: the first goal is to increase the accuracy of classification and the second goal is to reduce the number of fuzzy sets for each feature. The ABC algorithm tries to find the best segmentation for the features by creating several fuzzy sets in the form of three worker bees, observer and discoverer and performing various searches with the two goals of accuracy and reduction of complexity. The results show that the proposed ABC-FDT method has the best prediction accuracy with 96.87% compared to other methods. In future research, it is suggested that in addition to setting the number of fuzzy sets for each property, the decision threshold values be optimized. Determining the optimal decision values can reduce the number of fuzzy sets for properties and ultimately the number of rules.

REFERENCES

[1] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. *Computational and structural biotechnology journal*, 15, 104-116.

[2] Rezaeipanah, A., Mojarad, M., & Sechin Matoori, S. (2021). Intrusion Detection in Computer Networks Through Combining Particle Swarm Optimization and Decision Tree Algorithms. *Journal of Business Data Science Research*, 1(1), 14-22.

[3] Rezaeipanah, A., & Ahmadi, G. (2020). Breast Cancer Diagnosis Using Multi-Stage Weight Adjustment In The MLP Neural Network. *The Computer Journal*.

[4] Selvakumar, S., Kannan, K. S., & GothaiNachiyar, S. (2017). Prediction of diabetes diagnosis using classification based data mining techniques. *International Journal of Statistics and Systems*, 12(2), 183-188.

[5] Ghalehgolabi, M., & Rezaeipanah, A. (2017). Intrusion detection system using genetic algorithm and data mining techniques based on the reduction. *International Journal of Computer Applications Technology and Research*, 6(11), 461-466.

[6] Jain, V., & Raheja, S. (2015). Improving the prediction rate of diabetes using fuzzy expert system. *IJ Information Technology and Computer Science*, 7(10), 84-91.

[7] Choubey, D. K., Paul, S., Kumar, S., & Kumar, S. (2017, February). Classification of Pima Indian diabetes dataset using naive bayes with genetic algorithm as an attribute selection. In *Communication and Computing Systems: Proceedings of the International Conference on Communication and Computing System* (pp. 451-455).

[8] Pradhan, M. A., Bamnote, G. R., Tribhuvan, V., Jadhav, K., Chabukswar, V., & Dhobale, V. (2012). A genetic programming approach for detection of diabetes. *Int J Comput Eng Res (ijceronline. com)*, 2(6), 91.

[9] Saybani, M. R., Shamshirband, S., Golzari, S., Wah, T. Y., Saeed, A., Kiah, M. L. M., & Balas, V. E. (2016). RAIRS2 a new expert system for diagnosing tuberculosis with real-world tournament selection mechanism inside artificial immune recognition system. *Medical & biological engineering & computing*, 54(2-3), 385-399.

[10] 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.

[11] Ahmad, F., Isa, N. A. M., Hussain, Z., Osman, M. K., & Sulaiman, S. N. (2015). A GA-based feature selection and parameter optimization of an ANN in diagnosing breast cancer. *Pattern Analysis and Applications*, 18(4), 861-870.

[12] Pourpanah, F., Lim, C. P., & Saleh, J. M. (2016). A hybrid model of fuzzy ARTMAP and genetic algorithm for data

classification and rule extraction. *Expert Systems with Applications*, 49, 74-85.

[13] Cheruku, R., Edla, D. R., & Kuppili, V. (2017). SM-RuleMiner: Spider monkey based rule miner using novel fitness function for diabetes classification. *Computers in biology and medicine*, 81, 79-92.

AUTHORS PROFILE

Mr. Mousa Mojarad received his PhD in Computer-Software Engineering in 2020. He is currently a lecturer and faculty member of the Islamic Azad University, Firoozabad Branch. His hobbies are big data, cognitive computing, clustering, software engineering, classification models, and cloud computing. He has more than 8 years of teaching experience and 6 years of research experience.

Mr. Esmaeil Hajizadegan received the B.Sc. Degree in *Misagh Institute* of Higher Education, Rafsanjan, Iran in 2011, and the M.Sc. degree in computer software engineering from Liyan Institute of Education, Bushehr, in 2021. His primary research interest is in using supervised learning algorithms for routing, although he has concurrent research in artificial neural network, algorithms, computer architecture, computer engineering, and communication engineering.

Mr. Mohsen Gurkani received the B.Sc. Degree in Islamic Azad University, Bafgh Branch, Iran in 2010, and the M.Sc. degree in computer software engineering from Liyan Institute of Education, Bushehr, in 2021. Mohsen's current research interests include the machine learning, classification, pattern recognition, supervised learning, information technology, optimization, and feature extraction.

